UNIVERSIDADE DE LISBOA
FACULDADE DE LETRAS

A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students

Tanara Zingano Kuhn

Orientadores: Prof.ª Doutora Margarita Maria Correia Ferreira
Prof. Doutor Carlos Alberto Marques Gouveia

Tese especialmente elaborada para obtenção do grau de Doutor no ramo de Linguística, na especialidade de Linguística Aplicada

2017

UNIVERSIDADE DE LISBOA
FACULDADE DE LETRAS

A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students

Tanara Zingano Kuhn

Tese especialmente elaborada para obtenção do grau de Doutor no ramo de Linguística, na especialidade de Linguística Aplicada

Orientadores: Prof.ª Doutora Margarita Maria Correia Ferreira
Prof. Doutor Carlos Alberto Marques Gouveia

Júri:

Presidente: Prof.ª Doutora Maria Inês Pedrosa da Silva Duarte, Professora Catedrática e Membro do Conselho Científico da Faculdade de Letras da Universidade de Lisboa

Vogais:

- Doutora Ana Lúcia Frankenberg-Garcia, Reader School of Literature and Languages, University of Surrey, U.K.;
- Doutor Iztok Kosem, Assistant with a Ph. D Faculty of Arts, University of Ljubljana, Slovenia;
- Doutora Maria Joana de Almeida Vieira Santos, Professora Auxiliar Faculdade de Letras da Universidade de Coimbra;
- Doutora Margarita Maria Correia Ferreira, Professora Auxiliar Faculdade de Letras da Universidade de Lisboa, orientadora;
- Doutora Maria Amália Pereira Mendes, Investigadora Auxiliar Centro de Linguística da Faculdade de Letras da Universidade de Lisboa;
- Doutora Sandra Maria Brito Pereira, Bolseira de Pós-Doutoramento (FCT) Centro de Linguística da Faculdade de Letras da Universidade de Lisboa.

Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (Capes)
Processo número 0973/13-0

2017

Acknowledgments

I would like to thank my supervisor, Margarita Correia, for all her guidance, support and belief in me. In many difficult moments, she would give me a word of comfort and reassurance. Her encouragement was essential in this moment of my life; without her supervision and constant help, this thesis would not have been possible. I cannot thank her enough for everything that she has done for me.

I wish to thank my co-supervisor, Carlos Gouveia, for our talks, especially for some apparently simple questions that would have me thinking for a long time. His course on academic writing contributed greatly to enriching my thesis. Moreover, his kindness and encouragement were very important in this phase of my life.

I would like to express my most profound gratitude to Iztok Kosem. First and foremost, for receiving me at the University of Ljubljana for a Short-term Scientific Mission (ENeL/COST Action), during which I had the unparalleled opportunity to learn about the state-of-the-art methods, tools, and resources for e-lexicography. Moreover, his attention, true interest in my project, and valuable advice contributed greatly to the development of my PhD research. Undoubtedly, this was a crucial moment in my PhD and I am truly thankful for everything that he has done for me.

Special thanks are due to José Pedro Ferreira for helping me with the computational part of the compilation of my corpus, for our enlightening conversations on corpora design and for always making himself available in case I needed help.

I would like to thank Maria José Bocorny Finatto for introducing me to lexicography and making me fall in love with it. She first gave me the exciting opportunity to take part in her research and later co-supervised me when I was in the Netherlands. I am also deeply indebted to her for her insightful suggestions for my PhD project. Maria José has always been an inspiration and I cannot thank her enough for everything that she has done for me.

Another fundamental person in my transitioning into lexicography was Paul Bogaards (*in memoriam*). His careful attention and interest in my project, together with his constant encouragement, were fundamental for my development in the area. Paul was

not only an excellent supervisor but also an extremely kind person, who was always concerned about my well-being in the Netherlands. It was a great honour and a privilege to have been close to Paul.

I am also deeply indebted to Robert Lew for his generosity. Besides kindly making his work available to me and patiently answering my (many) questions, he thoughtfully informed me about the European Network of e-Lexicography when it had just been launched. Given that this network gave me the opportunity to go to Slovenia for the scientific mission with Iztok, I cannot thank Robert enough for ultimately having made all this possible.

I wish to present my warmest thanks to a very special person: Valdir do Nascimento Flores. Words are not enough to express how grateful I am for having had the invaluable opportunity to do research with him for so many years, and how lucky I am for having him as a friend. He has always been a role model for me and I would not be who I am if our lives had not crossed.

I would also like to thank my dear friend Letícia Soares Bortolini for always being there for me. We have experienced a lot together and this time was no different. Her unconditional support and encouragement were essential for me to go through this challenging time of my life.

I am also very thankful to my dear friend Yulya Mysyuk for all her kindness and care. She always made sure that I had some breaks and arranged the nicest things for us to do. Our talks were (and are) very dear to me – a conversation with her and stress would become much more manageable.

I am truly thankful to my husband Miguel, who has supported me unconditionally during this entire period. His attention and encouragement were fundamental for making me keep going.

I would also like to thank my parents, Egídio and Elisabeth, and my sister, Ananda, for their full support in everything I do. It was very reassuring to know they were always cheering for me.

I also thank my mother-in-law, Tila, and father-in-law, Zé António, for their attention, support and care.

Special thanks go to Andrew Swearingen for proofreading the manuscript.

Last but not least, I would like to express my sincere appreciation to the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (Capes-Brazil) for the PhD scholarship, to CELGA-ILTEC (University of Coimbra) for funding essential elements of my PhD research, such as scripts, to the European Cooperation for Science and Technology (COST) through the European Network of e-Lexicography (ENeL) Action for a Short-Term Scientific Mission, and to the University of Ljubljana for granting me licence to use the tools required for developing my PhD research.

Abstract

University students are expected to read and write academic texts as part of typical literacy practices in higher education settings. Hyland (2009, p. viii-ix) states that meeting these literacy demands involves "learning to use language in new ways". In order to support the mastery of written academic Portuguese, the primary aim of this PhD research was to propose a design of an online corpus-driven dictionary of Portuguese for university students (DOPU) attending Portuguese-medium institutions, speakers of Brazilian Portuguese (BP) and European Portuguese (EP), either as a mother tongue or as an additional language.

The semi-automated approach to dictionary-making (Gantar et al., 2016), which is the latest method for dictionary compilation and had never been employed for Portuguese, was tested as a means of provision of lexical content that would serve as a basis for compiling entries of DOPU. It consists of automatic extraction of data from the corpus and import into dictionary writing system, where lexicographers then analyse, validate and edit the information. Thus, evaluation of this method for designing DOPU was a secondary goal of this research.

The procedure was performed on the Sketch Engine (Kilgarriff et al., 2004) corpus tool and the dictionary writing system used was iLex (Erlandsen, 2010). A number of new resources and tools were created especially for the extraction, given the unsuitability of the existing ones. These were: a 40 million-word corpus of academic texts (CoPEP), balanced between BP and EP and covering six areas of knowledge, a sketch grammar, and GDEX configurations for academic Portuguese.

Evaluation of the adoption of the semi-automated approach in the context of the DOPU design indicated that although further development of these brand-new resources and tools, as well as the procedure itself, would greatly contribute to increasing the quality of DOPU's lexical content, the extracted data can already be used as a basis for entry writing. The positive results of the experiment also suggest that this approach should be highly beneficial to other lexicographic projects of Portuguese as well.

Keywords: academic Portuguese, automated lexicography, corpus, dictionary, tools development

Resumo

No ensino superior, espera-se que estudantes participem, em maior ou menor extensão, em atividades de leitura e escrita de textos que tipicamente circulam no contexto universitário, como artigos, livros, exames, ensaios, monografias, projetos, trabalhos de conclusão de curso, dissertações, teses, entre outros. Contudo, essas práticas costumam se apresentar como verdadeiros desafios aos alunos, que não estão familiarizados com esses novos gêneros discursivos. Conforme Hyland (2009, p. viii-ix), a condição para se ter sucesso nessas práticas é "aprender a usar a língua de novas maneiras".

A linguagem acadêmica é objeto de pesquisa há muitos anos, sendo especialmente desenvolvida no âmbito da língua inglesa. Se por um lado, durante um longo período todas as atenções estavam voltadas para o *English for Academic Purposes* (EAP) (inglês para fins acadêmicos), tendo em vista o incomparável apelo comercial dessa área, mais recentemente tem-se entendido que falantes de inglês como língua materna também precisam aprender inglês acadêmico, pois, como dito acima, trata-se de uma nova maneira de usar a língua, que os estudantes universitários desconhecem. Nesse sentido, é natural que a grande maioria de matérias pedagógicos como livros, manuais, gramáticas, listas de palavras e dicionários, por exemplo, sejam produzidos para o contexto de uso da língua inglesa.

Assim como o inglês e tantas outras línguas, o português também é usado em universidades como língua na e pela qual se constrói conhecimento. Aliás, nos últimos 15 anos, temos vivenciado um fenômeno de expansão do acesso ao ensino universitário no Brasil, paralelamente a um grande aumento da presença de alunos estrangeiros fazendo ensino superior no Brasil e em Portugal, o que reforça a natureza do português como língua de construção e difusão científica. É de se saudar os esforços e as medidas de política linguística da Comunidade dos Países de Língua Portuguesa (CPLP) para apoiar e fomentar o português como língua da ciência.

Apesar dessa clara importância do português acadêmico, sabemos que sua presença como objeto de estudo de uma área específica ainda é bastante restrita. Tem-se observado algum crescimento no que diz respeito à abordagem discursiva da linguagem acadêmica; contudo, descrições ao nível léxico-gramatical ainda são bastante escassas.

Em especial, no que concerne recursos lexicográficos como auxiliares pedagógicos, a existência de um dicionário de português acadêmico especialmente criado para atender as necessidades de estudantes universitários é desconhecida.

Nesse sentido, tendo em vista a demanda apresentada acima e a lacuna nos estudos atuais, a presente pesquisa de doutorado buscou colaborar tanto com o campo dos recursos ao ensino de português acadêmico quanto com o de elaboração de recursos lexicográficos através da proposta de desenho de um dicionário *online corpus-driven* de português para estudantes universitários (DOPU). Baseando-se em uma perspectiva de português como língua pluricêntrica, este dicionário contempla as variedades português brasileiro (PB) e europeu (PE). Além disso, o público-alvo se constitui por falantes de português como língua materna e como língua adicional.

Para a construção do desenho, adotou-se a mais moderna abordagem de compilação de dicionários atualmente existente, qual seja, a *semi-automated approach to dictionary-making* (Gantar et al., 2016). Esse método consiste na extração automática de dados de um corpus e importação para um sistema de escrita de dicionários, no qual lexicógrafos analisam, editam e validam as informações que foram automaticamente pré-organizadas nos campos da entrada conforme definições previamente estabelecidas. Esta abordagem é revolucionária no sentido em que o ponto de partida da análise lexical do *corpus* não mais se dá na ferramenta de análise de *corpus*, mas sim diretamente no sistema de escrita de dicionários. Experimentar essa abordagem no desenvolvimento do desenho do DOPU constitui-se em um objetivo secundário desta pesquisa de doutorado, uma vez que tal método nunca foi aplicado para a construção de dicionários de português.

Os programas utilizados para a aplicação do procedimento de extração foram o *Sketch Engine* (SkE) (Kilgarriff et al., 2004), provavelmente a mais sofisticada ferramenta de criação, análise e manutenção de *corpus* da atualidade, e o iLex (Erlandsen, 2010), um sistema de escrita de dicionários bastante flexível e com alta capacidade de processamento de dados.

Para a implementação da abordagem, são necessários: um *corpus* anotado com classes de palavra; uma *sketch grammar* (trata-se de um arquivo com relações gramaticais e diretivas de processamento para o sistema do SkE computar diferentes

tipos de relações através de cálculos estáticos); uma configuração de GDEX, isto é, *Good Dictionary Examples* – bons exemplos para dicionários (trata-se de uma configuração com classificadores para avaliar frases e atribuir pontuações conforme os critérios estabelecidos); e definições de parâmetros (frequência mínima dos colocados e das relações gramaticais). Tendo em vista a inadequação de *corpora* de português, bem como da *sketch grammar* e do GDEX existentes para o português, em função do propósito dessa extração de dados, qual seja, a compilação de entradas para o DOPU, foi necessário elaborar novos recursos.

Foi compilado o *Corpus de Português Escrito em Periódicos* (CoPEP), com 40 milhões de palavras, equilibrado entre as variedades PB e PE, e que cobre seis áreas de conhecimento. Os metadados do corpus foram detalhadamente anotados, permitindo fazer pesquisas avançadas. É o primeiro corpus internacional de português acadêmico de que temos notícia. De forma a padronizar a análise lexical e diminuir desequilíbrios na contagem estatística, o CoPEP foi pós-processado com o conversor Lince de forma a atualizar as ortografias de cada variedade conforme a determinação do Acordo Ortográfico da Língua Portuguesa, de 1990.

Uma *sketch grammar* foi especialmente elaborada para o CoPEP, e, nesse sentido, pode ser aplicada a outros *corpora* de português anotados pelo mesmo anotador. Optou-se por usar o anotador oferecido por padrão no SkE, qual seja, o *Freeling v3*. Criou-se uma *sketch grammar* com mais e mais precisas relações gramaticais do que aquela oferecida por padrão pelo SkE. Assim, usuários trabalhando com *corpora* de português anotados com *Freeling* no SkE poderão usar a minha versão, que já está disponível no *Sketch Engine*.

Uma configuração de GDEX havia sido produzida para fornecer exemplos para a compilação do *Oxford Portuguese Dicionary* (2015). No entanto, por ser bastante geral, elaborada para um *corpus Web* e por buscar selecionar exemplos para um dicionário bilíngue português-inglês/inglês-português, julgou-se mais apropriado criar uma configuração completamente nova. Assim, desenvolvi tal recurso, tendo em vista as características de uso da língua como apresentadas no CoPEP e o perfil do usuário do DOPU.

O procedimento de extração automática de dados do CoPEP e importação para o iLex tomou como base o procedimento usado para a criação de dicionários de esloveno (criadores desse método), fazendo-se adaptações. Acrescentaram-se dois elementos ao processo de extração: o *longest-commonest match* (LCM), que mostra a realização mais comum do par *keyword* e colocado, ajudando a entender o uso mais típico das colocações; e sugestões para atribuição de etiquetas com variedade típica, tanto para a *keyword* quanto para o colocado.

A avaliação do processo de escrita de entradas-piloto indicou que o método de extração de dados do CoPEP e importação para o iLex foi extremamente positivo, dado que a análise lexical pôde ser bastante sofisticada sem demandar o tempo rotineiro necessário quando se parte das linhas de concordância para elaboração de entradas. Alguns dados que nesta pesquisa não foram extraídos automaticamente e que tiveram que ser analisados manualmente na ferramenta de *corpus* poderão ser incluídos numa próxima versão do procedimento. Análise do processo de criação dos recursos necessários indicou que aprimoramentos podem ser feitos, assim aumentando a acurácia da extração.

Espera-se que o desenho de dicionário *online corpus-driven* de português para estudantes universitários proposto por esta pesquisa de doutorado sirva como base para o desenvolvimento de outras pesquisas relacionadas de forma que a sustentar a elaboração do DOPU.

Palavras-chave: corpus, desenvolvimento de recursos, dicionários, lexicografia automatizada, português acadêmico

Contents

List of Figures

List of Tables

List of Abbreviations

AO45 – *Acordo Ortográfico da Língua Portuguesa de 1945* (‘The Portuguese Language Orthographic Agreement of 1945’)

AO90 – *Acordo Ortográfico da Língua Portuguesa de 1990* (‘The Portuguese Language Orthographic Agreement of 1990’)

BP – Brazilian Portuguese

Capes – *Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior* (‘Coordination for the Improvement of Higher Education Personnel’)

Celpe-Bras – *Certificação de Proficiência em Língua Portuguesa para Estrangeiros* (‘Certificate of Proficiency in Portuguese for Foreigners’)

CL – Corpus Linguistics

CoPEP – *Corpus de Português Escrito em Periódicos* (‘Corpus of Portuguese from Academic Journals’)

COST – European Cooperation in Science and Technology

CPLP – *Comunidade dos Países de Língua Portuguesa* (‘Community of Portuguese-speaking Countries’)

CQL – Corpus Query Language

CQS – Corpus Query System

CUTe – Corpus of Undergraduate Students

DGEEC- *Direção Geral de Estatísticas da Educação e Ciência* (‘Directorate-General for Education and Science Statistics’)

DOPU – *Dicionário online de português para estudantes universitários* (‘Online Dictionary of Portuguese for university students’)

DTD – Dictionary Type Documentation

DWS – Dictionary Writing System

EAP – English for Academic Purposes

ENeL – European Network of e-Lexicography

EP – European Portuguese

ESP – English for Specific Purposes

FO43 – *Formulário Ortográfico de 1943* (‘The Orthographic Reform of 1943’)

GDEX – Good Dictionary Examples

HEI – Higher Education Institution

IILP – *Instituto Internacional da Língua Portuguesa* (‘International Institute of the Portuguese Language’)

IsF – *Idiomas sem Fronteiras* (‘Language without Borders’)

LCM – Longest Commonest Match

PAL – Portuguese as an Additional Language

PALis – *Plano de Ação de Lisboa* (‘Action Plan of Lisbon’)

POS – Part of speech

SciELO – Scientific Electronic Library Online

SFL – Systemic-Functional Linguistics

SkE – Sketch Engine

SLA – Second Language Acquisition

VOC – *Vocabulário Ortográfico Comum da Língua Portuguesa* (‘Common Orthographic Vocabulary of the Portuguese Language’)

Introduction

University students are expected to read and write academic texts as part of typical literacy practices in higher education settings. The magnitude of these tasks should not be underestimated, as one of the conditions for students to engage in these routine activities involves "learning to use language in new ways" (Hyland, 2009, p. viii-ix).

Such particular way of language use in academic (oral or written) texts has been traditionally called *academic language* and has its own characteristics, as demonstrated by a number of studies based on various theoretical frameworks (e.g. Biber, 1996, 2006; Biber & Conrad, 2009; Biber, Johansson, Leech, Conrad, & Finegan,1999; Hyland, 2009; Hyland & Bondi, 2006; Swales, 1990; to name but a few). It seems sensible to conclude that university students need **to learn** this new language in order to participate in higher education literacy activities.

Indeed, Swales (1990) calls for the explicit teaching of academic language. Nonetheless, the appeal most often referred to foreign students. Hyland (2006, p.2) shows that this view was later broadened, with the acknowledgement that academic language (in this case, English) must be taught to university students irrespective of their language status background:

> (…) there is growing awareness that students, including native English-speakers, have to take on new roles and engage with knowledge in new ways when they enter university. They find that they need to write and read unfamiliar genres and participate in novel speech events.

The recognition of the importance of teaching/learning academic language resulted in a number of studies and the development of pedagogical material by researchers of many languages. Among them, it is not surprising that the most attention has been given to English due to its dominance as the academic lingua franca around the world. Therefore, teachers and students dealing with academic English have at their disposal a plethora of textbooks, manuals, word lists, grammars, and dictionaries. Nevertheless, other languages also need resources for teaching and learning their own academic languages. One such language is Portuguese.

This call is further supported by the fact that currently, there is a very large community of users of academic Portuguese all over the world. Such high numbers are a result of at least two well-known phenomena: expansion of access to higher education in Brazil (in the 2000s) and Portugal (in the 1990s); and the internationalization of Portuguese. These university students are: speakers of Portuguese as a mother tongue studying at universities in Brazil, Portugal, Angola, for example, and speakers of Portuguese as an additional language (henceforth PAL) studying either at universities in Brazil, Portugal, Mozambique, etc., or in non-Portuguese speaking countries where Portuguese is offered as a major/minor or graduate course.

One of the consequences in the context of pedagogical practices is that the use of academic Portuguese can no longer be taken for granted. After all, the academic routine of this massive population of university students (more than 7 million people in Brazil and Portugal together, according to recent censuses) involves activities like reading and/or writing exams, monographs, articles, textbooks, abstracts, MA dissertations, among others, **in Portuguese.**

Taken together, the points raised above clearly suggest that research on academic Portuguese to support the production of pedagogical materials is of utmost importance. On a positive note, it seems that this call is starting to be heard, as an increased interest in literacy practices in higher education has been recently observed (cf. Carvalho, 2013; Cristóvão, Bork &Vieira, 2015), along with the first attempts at descriptions of academic Portuguese (cf. Molsing & Perna, 2014; Nunes & Perna, 2015). Nevertheless, research on academic Portuguese producing a solid body of studies is still rather incipient in Brazil and Portugal. This means research-informed teaching and learning resources for academic literacy in Portuguese are scarce. Among resources that students studying in Portuguese are lacking is a dictionary of academic Portuguese.

In view of this scarcity, the aim of my PhD research is to help to bridge this gap by proposing a design of a corpus-driven online dictionary of Portuguese for university students (henceforth DOPU). The target users of the proposed dictionary are tertiary level students, speakers of both Brazilian and European Portuguese, either as a mother-tongue or as an additional language.

Drawing on Kosem (2010), I claim that a fully-equipped dictionary of academic Portuguese should: 1) account for the characteristics of the Portuguese language used in academic texts; and 2) attend to university students' needs – independently of their language status (background and variety) or discipline of study. This call for a special dictionary is in line with Correia (2008, s.p.):

> Current good-quality dictionaries represent clearly delimited stretches of a language's lexicon and clear-cut frequency-based vocabulary sets according to the interests of dictionaries' target-users.[1] (my translation).

In order to comply with the above-mentioned criteria, it was decided to follow the corpus-driven approach, by which the corpus is used as the source of all information in the dictionary, from the headword list to the content of the entry. Concerning the need to accommodate different user profiles in terms of the physical tool, DOPU can benefit the most from customizability as a digitally-born dictionary created from scratch. That means that display of content can be adapted to meet each user's needs.

A key part of the proposal of DOPU design was the adoption of the semi-automated approach to dictionary-making, as originally proposed by Rundell and Kilgarriff (2011) and first implemented into lexicographic practice by Gantar, Kosem and Krek (2016). In this highly innovative approach, lexical data are automatically extracted from the corpus according to predetermined criteria and transferred to the dictionary writing system (henceforth DWS), where lexicographers then analyse, validate and edit the data to shape them into the final database entry. Such advanced methods stem from an ingenious combination of state-of-the-art technology in lexicography and computational linguistics.

So far, automatic extraction of data from the corpus and import into DWS has been successfully applied for making different types of dictionaries of Slovene (cf. Gorjanc, Gantar, Kosem, & Krek, 2017; Kosem, Gantar, Logar, & Krek, 2014; Kosem, Gantar, & Krek, 2013). As for dictionaries of Portuguese, to this author's knowledge,

[1] "*Os bons dicionários de hoje são representativos de fatias bem delimitadas do léxico de uma língua, de vocabulários claramente delimitados em função de critérios como a frequência de ocorrência das palavras e o seu interesse para o público-alvo visado pelo dicionário*".

no attempts have been made to employ this technique. For this reason, a significant part of this research is dedicated to the development of new resources and tools.

At this point, special attention should be drawn to the fact that throughout this thesis, the term *design* will be used as defined in Hartmann and James' *Dictionary of Lexicography* (1998):

> **design**. The overall principles that govern the production of efficient REFERENCE WORKS, taking into account not only features of content (INFORMATION CATEGORIES) and presentation (ARRANGEMENT), but also the reference needs and skills of the USER.

In other words, the reader should bear in mind that, rather than a prototype, the design of DOPU that is proposed here consists of an organised plan with a number of guidelines that have been determined through the association of theory with a hands-on approach.

The overall structure of the thesis takes the form of three Parts, followed by Discussions (Chapter 12) and Conclusions.

Part I – CALL FOR A CORPUS-DRIVEN DICTIONARY OF PORTUGUESE FOR UNIVERSITY STUDENTS – consists of a literature review, beginning with the presentation of further arguments for the compilation of a dictionary of Portuguese for university students (Chapter 1). Chapter 2 discusses the concept of academic language, providing a review of studies on central topics. I argue here for the adoption of the term "academic Portuguese" and present significant research that has been developed in Brazil and Portugal. Corpus and dictionary-making is the subject covered in Chapter 3. Firstly, corpus linguistics is reviewed. Special emphasis is given to the impact of the advances of computational technologies in lexicography. Thereafter, the relationship between corpora and Portuguese lexicography is accounted for, together with an evaluation of existing corpora of Portuguese with academic texts. Part I finishes with Chapter 4, where I set out a detailed plan for DOPU, presenting points of decision concerning fundamental aspects of dictionary-making that should structure the design.

Part II – SET-UP FOR SEMI-AUTOMATED LEXICOGRAPHY – presents the software, resources, and tools that were required for application of the semi-

automated approach to dictionary compilation. Chapter 5 introduces the Sketch Engine (Kilgarriff, Rychlý, Smrz, & Tugwell, 2004), which is probably the most sophisticated corpus tool currently available, showing in detail how the main functions that I used in this thesis work. Among them, the key function for the methodology applied in my PhD research was the word sketch– "a one-page, corpus-based summary of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2004, p.105); iLex (Erlandsen, 2010) was the dictionary writing system that was adopted and is also presented.

The unsuitability of existing corpora of Portuguese with academic texts for the purposes of my PhD research led me to compile a new corpus. Chapter 6 gives a comprehensive account of the process of compilation of the *Corpus de Português Escrito em Periódicos* - CoPEP ('Corpus of Portuguese from Academic Journals'), which contains 40 million words balanced between BP and EP, covering six areas of knowledge.

Chapter 7 describes in the detail the devising of a new sketch grammar for Portuguese developed especially for CoPEP. Sketch grammar is a file with grammatical relations and processing directives for the Sketch Engine system to compute different types of relations through statistical calculations. The data obtained from these computations then form the basis of the word sketch feature in the Sketch Engine, which is the heart of the process of automatic extraction of data from the corpus.

The last chapter of Part II concerns the development of GDEX configurations (Kilgarriff, Husák, Mcadam, Rundell, & Rychlý, 2008) especially for CoPEP for automatic examples selection. GDEX stands for Good Dictionary Examples and is a function in the Sketch Engine tool that, based on pre-defined criteria, identifies example sentences in the corpus, placing the best ones at the top of the list of concordance lines in order to facilitate the lexicographer's process of example selection.

Part III – THE DICTIONARY DESIGN PROPOSAL – comprises three chapters. Chapter 9 describes in detail the preparation for the procedure of automatic extraction of data from CoPEP and import into iLex, with a careful definition of steps. Two processes are described: the extraction experiment; and the second extraction, and the different phases are presented and explained. Chapter 10 accounts for the

macrostructure of DOPU, focussing on the definition of principles for the creation of candidate headword lists. Finally, the microstructure is covered in Chapter 11. Constituent elements of entries are presented and justification is given for such choices, including the illustration of some of the components.

Careful reflection on the development and results of this project is addressed in Chapter 12 – Discussions. I review the method of automatic data extraction and import into DWS, passing through each of the developed custom-tailored resources and tools, namely, CoPEP, Sketch Grammar, GDEX configurations and the extraction procedure. I also suggest enhancements to be applied to the design and highlight the contribution of my research to lexicography, linguistics, language teaching and linguistics policy. I end with the presentation of relevant pointers for future work.

In the Conclusions, I review some shortcomings of the design and propose venues for opening a debate about what to expect from future advancements in lexicography in general, and in lexicography of the Portuguese language in particular.

Part I

CALL FOR AN ONLINE CORPUS-DRIVEN DICTIONARY OF PORTUGUESE FOR UNIVERSITY STUDENTS

Introduction to Part I

The first part of this thesis provides arguments to support the call for an online corpus-driven dictionary of Portuguese for university students (DOPU). Moreover, it demonstrates that creating such a dictionary is entirely feasible.

I begin with the presentation of the socio-political context of such a call, encompassing key facts about the status of higher education in Brazil and Portugal, the recent phenomenon of the internationalization of the Portuguese language, the condition of Portuguese as a pluricentric language, and the role of the Community of Portuguese-speaking Countries (CPLP) in the context of linguistic policy. It becomes apparent that DOPU is not only a fundamental literacy tool, but also a political one.

In Chapter 2, I refer to a large body of studies on academic language to demonstrate that this register is different from other registers and that students are not familiar with it. Knowing how to use academic language is thus challenging and research-informed pedagogical material should be developed to help students in higher education. It is apparent that a series of issues concerning the lexicogrammatical level of academic Portuguese needs to be addressed, and the dictionary is the optimal tool to do so.

The history of the relationship between corpora and dictionary-making is addressed in Chapter 3 to show how it has evolved together with (or due to) computational technology to a point where dictionary creation is no longer a decades-long project. Access to state-of-the-art tools and resources has been proven to enable a streamlined lexicographical process. It is thus perfectly feasible to develop DOPU.

Part I finishes with Chapter 4, where a detailed plan of DOPU is described. All major factors contributing to the creation of DOPU are discussed, with a number of decisions made regarding the most fundamental parts of the project. In the end, this chapter provides a clear set of guidelines ready to be employed in a DOPU lexicographical project.

Chapter 1 On the need for a dictionary of academic Portuguese

This chapter calls for the creation of a dictionary of Portuguese targeted especially at university students. There are a number of reasons for this call, as will be shown in the next subsections. The topics that will be addressed include the phenomena of expansion of access to higher education in Brazil and Portugal and the internationalization of Portuguese. I also argue that the time is ripe for the development of a dictionary of academic Portuguese as a pedagogical resource for tertiary education because of the current repositioning of the Community of Portuguese-speaking Countries (CPLP) regarding matters of linguistic policy.

1.1 Expansion of access to higher education

Significant growth in higher education systems in Portugal and Brazil in recent decades has led to an increased number of students attending university (Almeida, Marinho-Araújo, Amaral, & Dias, 2012; Barros, 2015; Dias, 2015; Jezine, Chaves, & Cabrito, 2011).

Entry into tertiary education has grown exponentially in Brazil in the last 15 years. According to the Higher Education Census carried out annually by the *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira* (Inep[2]), in 2000[3] there were 2,695,927 enrolments; in 2015,[4] that figure jumped to 8,027,297– a growth of 297%.

This rapid growth is a direct consequence of political measures[5] which mainly focus on improving social conditions through the advancement of education.

[2] The National Institute for Educational Studies and Research "Anísio Teixeira" or Inep, is a special research agency linked to the Ministry of Education. For further information, see http://www.inep.gov.br.

[3] http://download.inep.gov.br/download/censo/2000/Superior/sinopse_superior-2000.pdf

[4] http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior

[5] Barros (2015) points out four main public political measures that have been taken in order to foster expansion and democratization of tertiary level education in Brazil: extension of financing private sector students through a scholarship programme (PROUNI) and bank loans (FIES); greater number of vacancies in public institutions through the opening of new educational establishments and the remodelling of existing ones, within the federal programme Reuni; incentives for distance learning education; and a policy of affirmative action quotas that establishes a 50% share of vacancies at public universities for students who attended public secondary school and come from low-income families.

Concerning future achievements, the National Plan for Education (PNE), which was passed by the Law Nº 13.005, from 25 June of 2014, aims at increasing that growth, alongside a series of other goals referring to all Brazilian educational levels, in the period of a decade – between 2014-2024. One of the objectives within *Meta 12* ('Aim 12') is to ensure the expansion of enrolment rate up to at least 40% in public higher education institutions.[6]

As it can be seen, this goal stems from the fact that, currently, expansion in tertiary education in Brazil has mostly happened by means of the private sector (Barros, 2015; Jezine et al., 2011; Nunes, 2007). Figure 1-1 shows the dramatic rise in the number of private institutions in a period of 15 years. In 2000, there were 1,180 Higher Education Institutions (HEI hereafter) in Brazil, of which 176 public and 1,004 private. Fifteen years later, of the total of 2,364 HEIs, 295 are public, while 2,609 are private. The private sector inflation becomes even more visible when the number of enrolments per sub-system is considered, as seen in Figure 1-2.

Within this group, vacancies are also calculated with regards to each state's percentage of black, mixed-race and indigenous population. Barros draws attention to the fact that, despite their good intentions, such measures have been controversial.

[6] In Brazil, public education is free. However, entry into public higher education institutions requires high level scores in national exams, favouring students with solid educational backgrounds who usually come from private primary and secondary schools, in detriment to students who lacked access to qualified basic education due to their attendance public primary schools, which, unfortunately, have suffered from neglect.

Unidade da Federação / Categoria Administrativa	2000		
	Total Geral		
	Total	Capital	Interior
Brasil	1.180	435	745
Pública	176	59	117
Federal	61	38	23
Estadual	61	21	40
Municipal	54	-	54
Privada	1.004	376	628

Unidade da Federação / Categoria Administrativa	2015		
	Total Geral		
	Total	Capital	Interior
Brasil	2.364	846	1.518
Pública	295	98	197
Federal	107	64	43
Estadual	120	33	87
Municipal	68	1	67
Privada	2.069	748	1.321

Figure 1-1 Number of Higher Education Institutions per location (capital and countryside) and administrative category in Brazil in 2000 and in 2015. (Inep[7])

Unidade da Federação / Categoria Administrativa			Total
Brasil			8.027.297
	Pública		1.952.145
		Federal	1.214.635
		Estadual	618.633
		Municipal	118.877
	Privada		6.075.152

Figure 1-2 Number of enrolments per type of administrative category (public, private) in Brazil, 2015 (Inep[8])

It has been argued that the phenomenon of privatization of higher education has advantages and drawbacks. On the one hand, private institutions have given opportunities to students who would not otherwise attend tertiary level institutions, like full-time workers with families, who are offered after working hours' courses, or students who lacked access to the same qualified basic education as students entering

[7] http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior

[8] http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior

public institutions (Britto, Silva, Castilho, & Abreu, 2008; Almeida, et al. 2012). On the other hand, the vast majority of private sector institutions have adopted a profit-driven business-oriented model, which favours profits at the expense of quality, thus providing limited educational conditions to students (Barros, 2015; Jezine et al., 2011).

In Portugal, while in 1974 only around 7% of the population attended tertiary level institutions, popular demand for equal educational opportunities following the Carnation Revolution (*Revolução dos Cravos*) and the requirement of alignment with European standards have resulted in a growth of 105.8% in 1990 (Almeida et al., 2012; Dias, 2015). The accommodation of such an impressive increase in attendance was possible due to the creation of public polytechnics and the opening of private institutions. The evolution of the number of enrolments in higher education in this period is represented in Figure 1-3.

Anos	Subsistema de ensino		
	Total	Público	Privado
— 1978	81.582	77.501	4.081
1979	79.436	72.830	6.606
— 1980	80.919	73.869	7.050
1981	83.754	75.658	8.096
1982	86.789	78.339	8.450
1983	89.310	80.382	8.928
1984	95.133	84.362	10.771
1985	102.145	88.416	13.729
1986	106.216	90.535	15.681
1987	117.128	94.280	22.848
1988	123.507	99.402	24.105
1989	135.937	107.033	28.904
— 1990	157.869	119.733	38.136

Fontes/Entidades: DGEEC/MEd - MCTES, PORDATA
Última actualização: 2016-10-04

Figure 1-3 Total number of enrolments and by education sub-system in Portugal from 1978 to 1990. (PorData[9])

[9]http://www.pordata.pt/Portugal/Alunos+matriculados+no+ensino+superior+total+e+por+subsistema+de+ensino-1017.

The steep rise in the number of the population of Portugal enrolled in tertiary education took place in the 1990's. At the beginning of the decade, there were 157,869 students in higher education. In 1999, enrolments reached the notable figure of 356,790.

Gradual growth continued at the beginning of the 21st century, with the number of enrolments reaching a peak (400,083) in 2003. A steady decline marked the following years until 2007, when this trend changed with participation in higher education increasing again. As of 2011, another shift took place, revealing a drop in the number of enrolments. In 2015, higher education attendance reached its lowest value since 1998. However, the promising slight increase seen in 2016 might suggest a new trend. The fluctuation in the number of enrolments in Portugal from 1999 to 2016 can be seen in more details in Figure 1-4.

Anos	Subsistema de ensino		
	Total	Público	Privado
1999	356.790	238.857	117.933
2000	373.745	255.008	118.737
2001	387.703	273.530	114.173
2002	396.601	284.789	111.812
2003	400.831	290.532	110.299
2004	395.063	288.309	106.754
2005	380.937	282.273	98.664
2006	367.312	275.521	91.791
2007	366.729	275.321	91.408
2008	376.917	284.333	92.584
2009	373.002	282.438	90.564
2010	383.627	293.828	89.799
2011	396.268	307.978	88.290
2012	390.273	311.574	78.699
2013	371.000	303.710	67.290
2014	362.200	301.654	60.546
2015	349.658	292.359	57.299
2016	356.399	297.884	58.515

Fontes/Entidades: DGEEC/MEd - MCTES, PORDATA
Última actualização: 2016-10-04

Figure 1-4 Total number of enrolments and by education sub-system in Portugal from 1999 to 2016. (PorData[10])

[10]http://www.pordata.pt/Portugal/Alunos+matriculados+no+ensino+superior+total+e+por+subsistema+de+ensino-1017.

Researchers indicate that such variance in the number of enrolments in the last 15 years is due to a number of factors. For instance, on the one hand, the reduction of birth rates, which contributed to a decrease in the number of the population enrolled in tertiary education. On the other hand, change in legislation that facilitated access to higher education for adults over 23 years old, together with an important reduction in dropouts from secondary school, resulted in an increase in the population attending higher education (Almeida et al., 2012; Dias, 2015).

Notwithstanding important context-based differences, the significant rise in the number of higher education students in both Brazil and Portugal is a result of the same phenomenon, that of "the 'massification' of higher education". This "is associated to greater democratisation of attendance by students from social and cultural backgrounds with fewer opportunities and less family tradition of academic education" (Almeida et al. 2012, p. 899).

From a pedagogical viewpoint, this massification reveals an increased demand for students' engagement in new literacy practices, namely, production and comprehension of texts such as essays, reports, articles, dissertations, theses, among others, **in Portuguese,** by an enormous university population.

If at any time in the past the justification for not investing in auxiliary resources of Portuguese for academic purposes, such as dictionaries, glossaries, and manuals, stemmed from an alleged "low demand", then at present, with approximately 8.5 million university students, Brazilian and Portuguese researchers can no longer ignore this usually forgotten sector of the education system.

Furthermore, in addition to the quantity factor, the demand for academic Portuguese material lies with the undeniable and crucial role of Portuguese in students' academic life. Acquisition of the academic knowledge that is required to complete a degree from first-year students to bachelors (and from bachelors to masters, to doctors, to post-doctors) takes place **through the use of Portuguese**.

Accordingly, I understand that it is our role as teachers, linguists and researchers of the Portuguese language, to do our best to guarantee that more students not only have access to a university education, but also that the education provided to them is highly qualified.

Taken together, the immense number of students in tertiary education who recognise the importance of Portuguese in their academic life indicates that there is a definite need for auxiliary learning resources. DOPU has thus been conceptualised and designed as a proposal for a specially tailored dictionary to meet the needs of higher education students.

1.2 Internationalization of Portuguese

One of the reasons for the current phenomenon of internationalization[11] of Portuguese is due to an increasing economic interest in Portuguese-speaking countries like Brazil and Angola. This new trend not only results in a significant growth in the broad area of Portuguese as an Additional Language (PAL), but also to an increased number of speakers of other languages pursuing their studies in universities where Portuguese is the medium of instruction, i.e. in CPLP[12] countries or countries where Portuguese is not the official language.

The substantial growth in the number of people applying for the Celpe-Bras exam[13] (127 applicants in its first edition, 1998; 10.946 applicants in 2016[14]) together with the inclusion of Portuguese as a Foreign Language in the governmental programme *Idiomas sem Fronteiras* (Languages without Borders)[15] clearly reflects this increasing interest in higher education in Brazil.

[11] The international media have profusely reported on this phenomenon. See, for instance, Monocles' special issue "Generation Lusophonia: why is Portuguese the new language of power and trade", 2012.

[12] The Community of Portuguese Language Countries 'is the privileged multilateral forum for deepening mutual friendship and cooperation among its members' (my translation) (http://www.cplp.org/id-2763.aspx), and comprises Angola, Brazil, Cape Verde, East Timor, Guinea-Bissau, Mozambique, São Tomé and Príncipe and Portugal. Equatorial Guinea became an official Member State of CPLP on July 23, 2014. In this country, Portuguese is one of the official languages, but is not spoken by the population. See section 1.4 for further information on CPLP.

[13] The only officially accepted Brazilian proficiency exam of the Portuguese Language. A certificate of Proficiency is a mandatory requirement to study at Brazilian universities. For further information on the Celpe-Bras exam, see http://portal.inep.gov.br/celpebras

[14] Information available at http:// http://www.ufrgs.br/acervocelpebras/estatisticas/numero-de-examinandos-homologados/view

[15] The *Idiomas sem Fronteiras* (IsF) programme was officially implemented through Decree Nº 973 of November 14th 2014 in *Diário da União (seção 1, p. 11)* and is a complement to the *Science without Borders Programme* and other governmental student mobility schemes. Primarily called *English without Borders*, hence focusing on English language learning by Brazilian university students, the IsF broadens the initial programme by including other languages, such as French and Portuguese. According to the Decree, one of the objectives of the IsF is (art. 2º) *"fortalecer o ensino de idiomas no país, incluindo o da*

As for Portugal, its long-existing tradition for the presence of international students in its academic institutions has been maintained. According to the *Direção Geral de Estatísticas da Educação e Ciência* (DGEEC)[16] ('Directorate-General for Statistics of Education and Science'), in 2013/14, 11. 687 foreign students were enrolled in tertiary level institutions from six months to one year (Credit International Mobility Programme), which represents 3.2% of the total number of enrolments in higher education in Portugal. The number of international students enrolled in Portuguese universities for full-time study was 14,883 (Graduation International Mobility Programme), which is 4% of the total number of enrolments.

In addition to the expanded university attendance of speakers of other languages in Brazil and Portugal, Portuguese instruction in institutions of higher education worldwide has increased. A part of this expansion is due to governmental actions like the readership programmes sponsored by Capes (Brazil) and *Camões I.C.L.* (Portugal); the other part results from initiatives by international universities'[17].

The internationalization of Portuguese and the consequent increase in the number of speakers of other languages seeking Portuguese-language academic settings reinforce the aforementioned call: that speakers of PAL also become target users of a dictionary of academic Portuguese.

These students face an even greater literacy challenge: to master academic language skills in a language other than their own. This understanding leads to the adoption of certain measures when creating DOPU in order to meet their special needs. The literature on English Learners' Dictionaries (e.g. Rundell, 1999) indicates, for

língua portuguesa, e, no exterior, o da língua portuguesa e da cultura Brasileira" ('to strengthen language teaching in the country, including that of the Portuguese language, and, abroad, that of the Portuguese language and Brazilian culture') (my translation). IsF's official website informs that "*suas ações também atendem a comunidades universitárias brasileiras que passam a receber um número cada vez maior de professores e alunos estrangeiros em seus campus*" ('its [the IsF's] actions also assist those universities which have been receiving an ever-growing number of foreign professors and students on their campuses') (my translation). For more information on the IsF Programme, see http://isf.mec.gov.br/.

[16] http://www.dgeec.mec.pt/np4/18/

[17] I was personally a part of such initiatives. In 2009, I was hired as a full-time lecturer of Portuguese at Hankuk University of Foreign Studies in Seoul, South Korea. The Portuguese and Brazilian Studies Department traditionally had two teachers of Portuguese, one from Brazil and the other from Portugal. It was the first time they had hired a second Brazilian teacher, due to students' increased interest in Brazil. In 2012, I taught as a part-time lecturer at Leiden University, in Leiden, the Netherlands, where a Major in Brazilian Studies, the only one in the whole country (at the time), had just been implemented in October 2011.

instance, that definitions should be written in a simple way, collocates should be highlighted, and frequency of occurrence of lexical items should be displayed[18].

In view of what has been stated previously, the fact that speakers of PAL are also university students needing to master academic Portuguese, reinforces the demand of a specially tailored dictionary of academic Portuguese. This augmentation is twofold in nature: quantitatively, these students increase the total number of university students using academic Portuguese; qualitatively, the dictionary should be especially concerned with their needs as speakers of PAL.

1.3 Portuguese as a Pluricentric language

Thus far, Brazilian and Portuguese higher education scenarios have been addressed. The justification for the necessity of a dictionary of academic Portuguese has been given based on information about the massive surge in tertiary education in these countries and the growing importance of the Portuguese language worldwide. Two questions arise from this: Why is there reference only to Brazil and Portugal, given that Portuguese is the official language in seven other countries? Moreover, why is only one dictionary of academic Portuguese proposed when it is known that Brazilian and Portuguese varieties present substantial differences?

The answers to these questions are connected. Firstly, the limited space of this PhD project does not allow for expansion[19] to the other Member States of CPLP (see section 1.4 below), which would require several additional actions. One would require a detailed investigation of higher education scenarios, the role of Portuguese in academic contexts, and state-of-the-art of academic language studies in each country. Another would involve the collection of written academic texts following strict design criteria (see Chapter 6 for details on the CoPEP compilation process). This enterprise would pose a daunting challenge due to the very few publications in Portuguese available (see

[18] Learner's dictionaries have adopted different systems to pass that information in an easy, explicit way. As an example, the Macmillan English Dictionary for Advanced Learners uses a red-star system, in which "three-star words are the most common 2,500 words in the language. Two-star words are the next most common, and one-star words are the next most common 2,500" (http://www.macmillandictionary.com/learn/red-words.html).

[19] This limitation is actually an interesting possible future work, as I shall demonstrate in the Conclusions.

Appendix A, Figures 13.1-13.3,[20] for statistics on publications per country and language on SciELO and Web of Science).

Moreover, it should be noted that in official discourse, Angola, Cape Verde, East Timor, Guinea-Bissau, Mozambique, and São Tomé and Príncipe[21] adopt the European variety as their standard language. It has been shown that some of these countries have uprising norms. Nevertheless, they are not yet fully described and codified (but see the case of VOC below). Thus, at present, description of the Brazilian and European varieties should potentially account for the needs of all higher education students of CPLP[22].

Secondly, by contemplating the Brazilian and Portuguese varieties in one lexicographical resource, I am endorsing a pluricentric-language viewpoint. This is not only the approach taken in this thesis from a theoretical standpoint but also a political position that I, as a researcher, assume and defend. According to Michael Clyne, editor of the seminal book *Pluricentric languages: Differing norms in different nations*, "the term *pluricentric* was employed by Kloss (1978 II, p. 66-67) to describe languages with several interacting centres, each providing a national variety with at least some of its own (codified) norms" (Clyne, 1992, p.1). The advent of pluricentricity eliminates the judgemental appreciation often given to language varieties: it does not make sense to state that a certain national variety is a "deviation from the centre", because there are many centres, and each one of them has its norms. This is the case of Portuguese.

Allan Baxter, in the same book, affirms that Portuguese is a pluricentric language existing in two standards, the Brazilian and European varieties and that "The two standards differ from each other in phonology, morphology, syntax, lexicon, spelling and pragmatics" (1992, p. 35). These differences do not entail, however, that a

[20] These figures are reproductions of tables from *Observatório da Língua Portuguesa*. Figure 13 1: https://observalinguaportuguesa.org/percentagem-de-publicacoes-dos-paises-da-cplp-por-idioma/;Figure 13 2: https://observalinguaportuguesa.org/publicacoes-cientificas-2/; Figure 13 3: https://observalinguaportuguesa.org/publicacoes-cientificas-web-of-science-e-scopus/

[21] Equatorial Guinea became an official Member State of CPLP on July 23, 2014. As mentioned earlier, Portuguese is one the official languages, but it is not spoken by the population.

[22] DOPU is planned to be an online under-construction dictionary, that is, "not 'a fixed object', but 'an organic, changing database" (Prinsloo 2001, p. 141, as cited in Klosa, 2013, p. 519). Thus, in the hope that this project is continued, it will be possible to include academic language from other Portuguese varieties in a future work.

dictionary of Portuguese should not cover both varieties, at least in an online dictionary, which enables a great deal of customization.

In fact, it should not be forgotten that such a strategy has been recently adopted for the creation of another lexical resource fostering Portuguese-as-a-pluricentric-language policy, the *Vocabulário Ortográfico Comum da Língua Portuguesa* ('Common Orthographic Vocabulary of the Portuguese Language') (VOC) (Ferreira, Correia & Almeida, 2017).[23] VOC is a freely accessible online platform that hosts the official regulatory orthographic vocabularies of the Member States of CPLP.

In 1990, ministers of the Member States of CPLP signed an agreement implementing an orthographic reform (*Acordo Ortográfico da Língua Portuguesa de 1990*, 'The Portuguese Language Orthographic Agreement of 1990', henceforth AO90 -) with the objective of the simplification and unification of spelling rules regulating Brazilian and European varieties of Portuguese (the other varieties adopt EP as their standard). At the time, there were two official spelling norms for Portuguese: the *Formulário Ortográfico de 1943* (henceforth FO43) ('The Orthographic Reform of 1943'), which was followed in Brazil, and the *Acordo Ortográfico de 1945* (henceforth AO45) ('The Portuguese Language Orthographic Agreement of 1945'), which was the norm in Portugal, Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Príncipe, and East Timor[24]. It is noteworthy that this spelling reform unifies the orthography of many words without suppressing national lexicography traditions. VOC was created to be the official international reference tool for regulation of national vocabularies and implementation of the spelling reform.

Nonetheless, the fairly troubled[25]circumstances of the actual execution of the Orthography Law postponed its official implementation to 2009. Still, during the period

[23] VOC was created by a multinational team of lexicographers, under the coordination of the International Institute of the Portuguese Language (IILP), which is the official bureau for language policy of CPLP. VOC is a state-of-the-art lexical resource that not only provides references to the official spelling of a word, but also to its complete inflectional paradigm and syllabic division. It is available at http://voc.cplp.org/

[24] For an evaluation of AO90 based on a retrospective review of orthographic reforms in the Portuguese language and with enlightening contextual descriptions, see Neves (2010).

[25] This refers to a number of events, from the unsuccessful first attempt of implementation (1990) to Brazil and Portugal's extension, in 2009, of a previously accorded transitional period of three years. It should be noted that these new rules have been opposed by many and generated a heated, passionate debate about "language corruption", "impoverishment of the Portuguese language", among other non-

of adjustment, pre-agreement and new agreement norms were officially accepted simultaneously. Following this deadline, AO90 has not yet been fully adopted in Portugal, with some people and media still using the old spelling.

Given the arguments raised above, the guiding principle in this thesis is as follows: Brazilian and European Portuguese are the standard varieties under study; there is no deviation from a central norm, but different norms exist for each of these varieties. Accordingly, the treatment given to the description of Portuguese in this work should be likewise: it is not possible to try to iron out the differences for the sake of a uniform and homogeneous language. On the contrary, by adopting this pluricentric view, the Portuguese language here scrutinised shall be fully described and explained according to its behaviour and use by its different speakers, as shall be demonstrated by the corpus analysis. Only when differences are not present is such a distinction unnecessary.

DOPU will be able to account for this pluricentric language viewpoint due to its customisable user search interface, which prompts the user to choose a variety.

1.4 Alignment with CPLP linguistic policy

The Community of Portuguese-speaking Countries (CPLP) 'is the privileged multilateral forum for deepening mutual friendship and cooperation among its members[26]'. One of its three goals is to foster promotion and dissemination of the Portuguese language. In order to help with that aim, thus strengthening the contacts between the Member States and their technical teams, CPLP has the support of the International Institute of the Portuguese Language[27] (IILP). The fundamental objectives of this institution are 'to foster, to preserve, to enrich and to disseminate the Portuguese language as a means of culture, education, information, access to scientific and technologic knowledge, and as an official language used in international forums'.[28]

---

sensical arguments, mostly in Portugal. In this country, it is common to see disclaimers at the end of published texts informing the author's position to not follow AO90 spelling.

[26] "*é o foro multilateral privilegiado para o aprofundamento da amizade mútua e da cooperação entre os seus membros*" (my translation), in https://www.cplp.org/id-2763.aspx

[27] http://iilp.cplp.org/iilp.html

[28] "*a promoção, a defesa, o enriquecimento e a difusão da língua portuguesa como veículo de cultura, educação, informação e acesso ao conhecimento científico, tecnológico e de utilização oficial em fóruns internacionais*" (my translation), in https://iilp.wordpress.com/about/

Based on an initiative of the Portuguese Government and with the coordination of CPLP, Camões I.C.L., IILP and a consortium of four Portuguese Universities (Universities of Lisbon, Porto, Coimbra and Nova de Lisboa), the *II International Conference about the Future of the Portuguese Language in the World System* took place from 29 to 31 October 2013, in Lisbon. It gathered researchers and teachers of all the Member States of CPLP, who presented papers and proposed discussions on six themes: Science and Innovation; Internationalization; Teaching and Teacher Education; Linguistic Diversity: policies; Orthographic Vocabularies; and Teaching and Development.

The conclusions drawn from this conference resulted in the Action Plan of Lisbon (PALis), which was approved by the Council of Ministers of CPLP on 20 February 2014. This plan puts forward suggestions regarding the following topics: i. Portuguese as a Language of Science and Innovation; ii. Portuguese as a relevant factor in the creative economy; iii. Portuguese in the cooperation among the Member States of CPLP and diaspora communities; iv. Portuguese in international organizations; v. Teaching Portuguese to speakers of other languages (PALis: 5).

The recommendations for topic *i – Portuguese as a Language of Science and Innovation* stem from the recognition that:

> As a language of science, Portuguese faces some challenges in the global context, with enormous qualitative requirements, in which Portuguese-speaking researchers still have a low rate of participation (my translation)[29](PALis: 6).

Based on this claim and on the strategic dimensions that CPLP endorses, which largely refer to cooperation in investigation, scientific knowledge as a means for social inclusion and poverty reduction, human resources development for integration of CPLP citizens in the international community, and technical-scientific cooperation as element of progress, PALis proposes that a series of seventeen measures be taken. Among them, I highlight three, which are directly related to this PhD thesis:

[29] "*Como língua de ciência, a língua portuguesa enfrenta alguns desafios no contexto mundial, de enorme exigência qualitativa, no qual a participação de investigadores de língua portuguesa é ainda escassa*".

> 1. To recommend that the Member States value, through support from their scientific institutions of financing, assessment and certification, the use of Portuguese in scientific communication and production, as well as in all working papers, applications and the management of scientific projects;
>
> 2. To map those scientific domains which, due to the nature of their objects of study and development spaces, contribute to the production of specialised scientific literature in Portuguese;
>
> 5. To promote the development of open access, free scientific resources and digital infrastructures in Portuguese[30](PALis: 7-8; my emphasis).

The dictionary of academic Portuguese that this PhD project seeks to design will be a fundamental auxiliary tool for the writing of papers, work documents and applications for scientific projects. Furthermore, due to its very nature, it shall help to promote the production of scientific literature. Finally, my proposal is that this dictionary constitute a free, online resource available for users all over the world who need help with academic Portuguese.

As previously stated, the target user of this dictionary is also a learner of Portuguese as an additional language. In that regard, my project aligns with strategic recommendation *v - Teaching the Portuguese language for speakers of other languages*. PALis acknowledges the phenomenon of the internationalization of Portuguese and the consequent increasing search for learning Portuguese in international universities (p. 14) and puts forward five initiatives to foster the teaching of PAL. My PhD research specifically contributes to the implementation of two of them:

> 1. To recommend that the IILP resume the discussion on strategies for Portuguese teaching which take the different varieties into account and continue with the creation of common teaching resources for Portuguese as a Foreign Language, as has been put into practice within the scope of the Portal of the Teacher of Portuguese as a Foreign Language (PPPLE);

[30]"*1. Recomendar aos EM que, por intermédio das suas instituições de financiamento, avaliação e certificação científicas, valorizem o uso da língua portuguesa na comunicação e produção científicas, assim como nos documentos de trabalho, de candidatura e gestão de projetos científicos; 2. Mapear os domínios científicos que, pela natureza do seu objeto e dos espaços em que se desenvolvem, propiciem a produção de literatura científica especializada em língua portuguesa; 5. Promover o desenvolvimento de recursos científicos e de infraestruturas digitais em língua portuguesa, de acesso aberto e gratuito*." (My translation).

> 3. To create and adjust auxiliary tools for the teaching of Portuguese for Specific Purposes[31](PALis: 15; my emphasis).

As can be seen, DOPU is in absolute consonance with PALis' recommendations. Firstly, it is a pluricentric language dictionary for university students. Secondly, it is a pedagogical resource that can support teaching and learning academic Portuguese in its two standard varieties. Thirdly, it is aimed not only at speakers of Portuguese as a mother tongue, but also speakers of PAL. Finally, as a dictionary of academic Portuguese, it is an auxiliary tool supporting a very specific area of teaching and learning Portuguese.

1.5 Summary

Taken together, the facts presented in the four previous sections, namely, expansion of access to higher education, internationalization of Portuguese, valorization of a pluricentric perspective, and alignment with linguistic policy put forward by CPLP[32], lead to the inevitable conclusion that the teaching and learning of academic Portuguese requires immediate support through, among other means, the development of pedagogical materials. A dictionary especially created to comply with university students' needs and with a pluricentric view of the Portuguese language is unheard of.

Given this urgent demand, I intend to contribute to addressing this serious gap with the proposal for a design of an online, corpus-driven dictionary of Portuguese for university students (DOPU).

[31] "*1. Recomendar ao IILP que retome a reflexão sobre estratégias de ensino da língua portuguesa que tenham em conta as suas diferentes variantes e que prossiga a criação de recursos didáticos comuns para o seu ensino como língua estrangeira, como tem sido realizado no âmbito do Portal do Professor de Português Língua Estrangeira (PPPLE); 3. Criar e aperfeiçoar instrumentos de apoio ao ensino da língua portuguesa para fins específicos*".

[32] At this point, I refer to Oliveira (2013), which is an instructive paper written by the former president of IILP on the change of position of CPLP and how it impacted the situation of Portuguese as an international language.

Chapter 2 Academic language

This chapter sets out to explain the definition of academic language underlying the present research. In section 2, a theoretical statement of my own view is provided. Then, a brief overview of some of the most frequent research topics of interest in the field specifically with regards to linguistic description is given in 2.1. Next, in 2.2, I argue for the use of *academic Portuguese*, justifying my position in light of the concept of academic language presented earlier. In 2.3, an overview of studies on academic Portuguese is presented, with a special focus on research on lexicogrammatical aspects. Some of these studies are summarised, with a focus on their outcomes and their contribution to designing DOPU. This chapter finishes with a discussion on academic vocabulary and the claim that DOPU shall be a fully-fledged dictionary, not a word list-cum-dictionary, as seen in section 2.4.

2.1 What is academic language?

University students are expected to read and write academic texts as part of typical literacy practices in higher education settings. Hyland states that meeting these literacy demands requires "learning to use language in new ways" (2009, p. viii-ix), in the sense that it differs from the ordinary use of language (cf. Hyland, 2006, 2009). Many studies have employed the terms *academic language*, *academic discourse*, and *academic prose* (most of the time interchangeably) to refer to this particular way of language use in higher education.[33]

In this thesis, what I consider to be academic language not only refers to the context of use, i.e. academia, but also to what Halliday and Martin (1996) define as *scientific language*, which is a functional variety of language, or register, in which:

> (…) certain words, and more significantly certain grammatical constructions, stand out as more highly favoured, while others correspondingly recede and become less highly favoured, than

[33] Numerous researchers adopt the term *academic* in reference to all levels of educational systems, i.e. elementary and secondary school and tertiary education. For an example related to the study of the Portuguese language, see the work by the "Discourse and Academic Discursive Practices" group at CELGA-ILTEC, under the coordination of prof. Carlos Gouveia.

> in other varieties of the language (Halliday & Martin, 1996, p. 4)

This concept was advanced within Systemic-Functional Linguistics (hereafter, SFL), in which register is seen as " a cluster of associated features having a greater-than-random (or rather, greater than predicted by their unconditioned probabilities) tendency to co-occur (Halliday, 1988 [Halliday & Martin, 1996, p. 59])". The typicality engendered by the co-occurrence of certain particular features is what makes it possible to distinguish language of science from, for instance, "journalese" (Halliday, 1990 [Halliday & Martin, 1996, p. 96]).

It should be stressed, however, that this understanding does not entail a view of scientific language as static or homogeneous. Given that *register*, as proposed by Martin (1992) based on Halliday's previous descriptions (cf. Halliday & Martin, 1996, p. 36), encompasses three variables, namely, field ('what is actually taking place' - the social action), mode ('what role is language playing' - the symbolic organization), and tenor ('who is taking part' - the role structure), then variation is intrinsic to scientific language. Halliday (1988 [Halliday & Martin, 1996, p. 59]) shed light on this topic with some examples of the terms in which variation takes place:

> in field, extending, transmitting or exploring knowledge in the **physical**, **biological** or **social sciences**; in tenor, addressed to specialists, to learners or to laymen, from within the same group (e.g., specialist to specialist) or across groups (e.g., lecturer to students); and in mode, phonic or graphic channel, most congruent (e.g., formal 'written language' with graphic channel) or less so (e.g., formal with phonic channel), and with variation in rhetorical function—expository, hortatory, polemic, imaginative and so on (my emphasis).

The emphasis on the excerpt above highlights that, although scientific 'language' is used in the singular, it is a general label that gathers different disciplines. This is especially relevant for grounding my definition of academic language.

An additional contribution from SFL to my argument here is the explanation for the nature of these distinctive features of scientific language. According to SFL rationale, these features simultaneously determine and are determined by scientific knowledge. Halliday and Martin further elucidate this point:

> The language of science is, by its nature, a language in which theories are constructed; its special features are exactly those which make theoretical discourse possible. But this clearly means that the language is not passively reflecting some pre-existing conceptual structure; on the contrary, it is actively engaged in bringing such structures into being. (Halliday & Martin, 1996, p. 9)

This dialectic relation between scientific language and science knowledge has been addressed with a grammatically-based approach in SFL. In this theory, grammar is understood as "the grammatical systems and structures — the clause complexes, clauses, phrases, groups and words, as well as the lexical items themselves —the vocabulary. So 'grammar' here is short for lexicogrammar" (Halliday, 1997 [2004, p. 183]). Moreover, it is treated as the realization of discourse (Halliday & Martin, 1996, p. 25); it is one stratum of the language. Language, as a meaning-making system (Halliday & Martin, 1996, p. 25), construes reality through grammar. Ultimately, grammar in scientific language construes a different reality than that construed by ordinary language:

> where the everyday 'mother tongue' of commonsense knowledge construes reality as a balanced tension between things and processes, the elaborated register of scientific knowledge reconstrues it as an edifice of things (Halliday & Martin, 1996, p. 17)

Key to this construal is a process called grammatical metaphor, which "is a substitution of one grammatical class, or one grammatical structure, by another; for example, *his departure* instead of *he departed*"(Halliday 1989 [2004 p. 87]). This example is only a simple illustration; it is not necessary, for the argument I am building here, to go into more complex structures. For now, what is relevant is to understand that this process involves objectification: for instance, processes and qualities, which in everyday language are expressed through verbs and adjectives, tend to be realised as nouns in scientific language. The conceptualization of grammatical metaphor[34] is thus

[34] Halliday points out grammatical metaphor as one of the features of scientific language that poses difficulties for students, in addition to interlocking definitions, technical taxonomies, special expressions, lexical density, syntactic ambiguity, and semantic discontinuity (Halliday, 1989 [Halliday & Martin, 1996, p. 76]).

one explanation for how and why language works differently when used for the production of scientific texts.

So far, I have demonstrated my understanding of academic language as a register of language, based on its characteristics and nature. However, one more concept needs to be invoked: that of genre.

One frequently quoted review of Genre Studies is Hyon 1996, in which she proposes the existence of "Three Traditions" reflecting different theories and resorting to illustrative and relevant works by researchers from each school: [35] (a) English for specific purposes (ESP) – e.g. Bhatia (1993); Flowerdew (1993); Swales (1990) (ibid, p. 698); (b) North American New Rhetoric studies – e.g. Bazerman (1988, 1994); Miller, (1984, 1994) (ibid, pp. 698-699) ; and (c) Australian systemic functional linguistics – e.g. Christie (1991, 1992); Martin (1989, 1991) (ibid., pp.700-701). In Brazil, one additional theoretical perspective has been added to these three traditions, the Socio-Discursive Interactionism (SDI), also known as the Geneva School (Bronckart, 1999; Schneuwly & Dolz, 1999) (cf. Bezerra, 2016, p. 467; Motta-Roth & Heberle, 2015, p.22).[36]

Notwithstanding epistemological differences among these four theoretical frameworks, I agree with the conclusions of Motta-Roth (2008, p. 350) and Bezerra (2016, quoting Motta-Roth, 2008) that these schools do have coinciding opinions about genres:[37]

1. Genres are uses of language associated with social activities;

[35] Hyon presents many more contributions and researchers; I here selected only a few.

[36] It should be mentioned that Brazil has its own version of Critical Genre Analysis (e.g., Bonini, 2002; 2010; Meurer, 2002, 2005; Motta-Roth, 1998; 2008a, as cited in Motta-Roth & Heberle, 2015), which is an interdisciplinary approach to genre studies characterised by utilizing concepts from ESP, SFL and Critical Discourse Analysis, has been adopted by many researchers. An example of an instructive and rich overview of the many theories of genre studies currently followed in Brazil is *Diálogos Brasileiros no Estudo de Gêneros Textuais/ Discursivos* (Lousada et al., 2016), with 89 selected contributions out of 750 communications presented at the 8th Symposium on Genre Studies, São Paulo, in 2015.

[37] Bhatia (2004, pp. 22-24) also states that different genre theories share common views on the conceptualization of genres. However, he only refers to the three traditions stated in Hyon 1996 and suggests five coinciding features.

2. These are recurrent discursive actions and thus have a certain degree of stability in form, content and style.[38] (my translation).

These two characteristics that transverse all four genre theories are in accordance with one of the criteria usually adopted for text classification in corpus linguistics. Given the confusion in naming such a criterion, with *genre*, *register*, *text type* tending to be employed indistinctively by many corpora compilers and investigators, Lee (2001), after extensive research on? these terms, creates an organised index for what he called "the BNC jungle",[39] adopting the term *genre*. In his understanding:

> Genre is used when we view the text as a member of a category: a culturally recognised artifact, a grouping of texts according to some conventionally recognised criteria, a grouping according to purposive goals, culturally defined (Lee, 2001, p.46).

It should be noted that Lee has drawn on the three genre traditions mentioned above to establish a definition more suitable for corpus linguistics purposes. His definition has given rise to a functional concept of genre that can be used to propose texts categories in corpus design and in corpus queries by lexicographers, teachers, researchers or students, offering corpus users an additional alternative to further restrict their searches, thus contributing to the examination of the language under study.

Having said that, in this thesis, academic language is understood as a register of language encompassing varied disciplines and used in context-situated social activities. (Oral or written) texts are relatively stable in form, content and style and are characterised by determined linguistic features that materialise the uncommonsense interpretation of reality required for construing scientific knowledge, thus diverging from other registers of language (ordinary,[40] journalistic, etc.).

---

[38] "*1. Gêneros são usos da linguagem associados a atividades sociais; 2. Essas ações discursivas são recorrentes e, por isso, têm algum grau de estabilidade na forma, no conteúdo e no estilo.*"

[39] The author proposed an index for categorising the 4,124 texts files (Lee, 2001, p. 37) in the British National Corpus (BNC). BNC was the first very large corpus at that time (compiled between 1991-1993), comprising 100 million tokens and originally labelled with very broad classifications. In order to allow users to qualitatively enhance their searches with more narrowly contextualised data analysis, Lee created a set of categories based on the concept of genre.

[40] In lexicography, "ordinary" is usually employed to refer to the use of language that is not somehow specialized; it is also called general language. I follow the lexicography tradition here as this is the scope of this thesis. However, I would like to note that my personal understanding of language necessarily implies use, i.e. it must be understood in terms of genre, in which language is always specialised: it is

As a result, academic language description in DOPU, even though necessarily narrowed to the lexicogrammar stratum of the language, acknowledges the interdependence between disciplines, registers, and genres.

2.2 Objects of study in academic language

As exposed above, the nature of academic language and its distinctive features pose serious challenges for students, becoming the object of studies by investigators from diverse theoretical paradigms (cf. Flowerdew, 2002; Hyland, 2006; Charles, Pecorari, & Hunston, 2009). However, the vast majority of them – and the most globally well-known - focus on academic English, which is not surprising, given the overwhelming dominance of academic English as a lingua franca (see Swales' compelling paper "English as *Tyrannosaurus Rex*", 1997).

In this subsection, I briefly review some of these studies as a means to indicate the most representative topics of interest regarding the lexicogrammatical description of academic language. Research on Portuguese has significantly increased in the last few years and is reviewed in a separate subsection (see 2.4 below).

One of the approaches widely adopted by investigators is corpus-based research. Applying this method, Biber (1996) found that linguistic association patterns (that is, lexical and grammatical associations) vary according to their interaction with non-linguistic patterns (that is, distribution across registers,[41] dialects and time) (Biber, 1996, p. 174), which means that:

> (…) linguistic association patterns are generally *not* valid for the language as a whole. Rather, linguistic and non-linguistic associations interact with one another, so that strong linguistic

always used to fulfil certain goals in a certain situated-context, with certain participants (projected or not). As a result, what is usually called "ordinary" language also involves specialization, for instance, small talk language, grocery shopping language, doctor's visit language, etc. Ferreira, Almeida and Correia (2013) share this view, arguing that, given the concept of genre, it is reasonable to question the status of "general language", suggesting the it is a hypothetical concept. Such an understanding is also in line with additional language teaching traditions, in which situated-language use is a well-known pedagogical framework (cf. Schlatter & Garcez, 2009).

[41]As mentioned in the previous subsection, *Register* and *genre* are two terms frequently used in academic language studies and are understood differently according to the theory in which they are employed, thus usually causing confusion to the reader (cf. Biber, 2006, pp.10-12; Lee, 2001; Sampson, 1997). For each author, the term originally employed is kept. For Biber, register refers "to situationally-defined varieties described for their characteristic lexicogrammatical features" (2006, p. 11).

> associations in one register often represent only weak associations in other registers. (Biber, 1996, pp.174-175)

Based on this principle, Biber, Johansson, Leech, Conrad and Finegan conducted a lengthy corpus-based investigation whose result is the monumental *Longman Grammar of Spoken and Written English* (1999), which "describes the range of grammatical features in English and compares the use of these features in four major registers: conversation, fiction, newspaper and academic prose" (Biber, 2009, p. 13). This grammar is a milestone in the area of academic language studies, since not only has it corroborated previous work indicating academic language as a disparate language variety, but also has provided a very comprehensive description of how and to what extent language varies across registers.

The well-established term *lexical bundles* was firstly employed in this grammar (cf. Biber & Barbieri, 2007, p. 264) to refer to frequent – usually not complete nor idiomatic – multi-word sequences (Biber, Conrad, & Cortes, 2004, p. 371) that form building blocks in discourse and whose functions are related to expressions of stance ("attitudes or assessments of certainty that frame some other proposition"), discourse organization ("relationships between prior and coming discourse") and referential framing ("make direct reference to physical or abstract entities, or to the textual context itself") (Biber et al., 2004, p.384).

The authors also stressed that multi-word sequences have been studied by many researchers, meaning that various frameworks identify and explain this phenomenon differently, mostly diverging in terms of idiomaticity. This multiple understanding is reflected in the variety of labels referring to multi-word sequences, as "'lexical phrases', 'formulas', 'routines', 'fixed expressions', 'prefabricated patterns' (or 'prefabs'), and 'lexical bundles'"(Biber et al., 2004, p. 372). A literature review has also shown the uses of 'chunks' 'clusters' and 'collocations'.[42] What seems to be unifying such diversity is a common understanding that lexical analysis should be based on a

[42] Given this multiplicity of terms, for matters of coherence the original terms used by each author will be maintained for the remainder of this thesis, notwithstanding the fact that sometimes more than one term is used in the same piece of research or that those terms might indicate diverging theoretical frameworks.

language-in-use principle, that is, according to the relationship words have with other words.[43]

Numerous studies have set out to examine the differences in academic language use between native and non-native speakers of English.[44] For instance, Ädel and Erman (2012) examined the use of 4-words bundles in English written productions of L1 speakers of Swedish and native speakers of English in linguistics, revealing less and different uses by non-native speakers. Another study is that of Dutra, Orfano and Sardinha (2014) who set out to contrast uses of stance bundles in native and multi-national (including Brazilian) non-native argumentative academic essays written in English. Focusing on the qualitative analysis of three subcategories, namely, hedge, epistemic stance and "obligations & directives" (Dutra et al., 2014, p.9), their findings revealed that native speakers tend to make use of more varied bundles for the same function, in contrast to the narrowed vocabulary employed by non-native speakers. Additionally, it has been shown that native speakers are less assertive, which is an expected writing characteristic in academic English, while non-native speakers tend to be much more direct and assertive towards their propositions.

Writing expertise is another dimension often examined in conjunction with discipline variation. Cortes (2004) identified and classified the functions (stance expression, discourse organization, and referential expression, as mentioned earlier) of

[43] In fact, as this author is concerned, single-item treatment of the lexicon appears to be the approach of choice only in the area of academic vocabulary research.

[44] The dichotomy native/non-native speaker has been widely used in different areas of linguistics, including Corpus Linguistics. Although this subject is beyond the scope of this thesis, as a linguist and a language teacher, I would like to state my position regarding what underlies such a dichotomy and elucidate the concept foregrounding my research and teaching. Firth and Wagner's seminal article (1997) called for a reconceptualization in the field of Second Language Acquisition (SLA), claiming that the concept of non-native speaker (among other fundamental ones), which is crucial to the cognitive, mainstream SLA approach, has given rise to "an analytical mindset that elevates an idealized "native" above a stereotypicalized "non-native", while viewing the latter as a defective communicator, limited by an underdeveloped communicative competence" (p. 285). Instead, they argued that attention should be drawn to participants (irrespective of their mother tongue) and how language learning takes place in social interaction. As a consequence, an analysis of successful communication events would replace an examination of errors and abnormalities, as the latter are concepts that epistemologically do not fit this new approach. This is not the place to review the important consequences of this paradigm shift for teaching-learning practices and language research (see Schlatter, Garcez & Scaramucci, 2004, for further discussion), however, I would like to stress that I follow the sociocultural SLA approach (see Bortolini&Kuhn, 2011). Thus, when the terms native/non-native (language, teacher, speaker, learner) are used in this thesis, they reflect the theoretical position of the author being mentioned, and not mine. For further information on the repercussions of Firth and Wagner's groundbreaking claim, see the focus issue of *The Modern Language Journal* "Second Language Acquisition Reconceptualized? The Impact of Firth and Wagner (1997)", 2007.

the most used lexical bundles in published articles in history and biology, then carried out verification of the use of these target bundles by students in three tertiary educational levels, concluding that students' use of target bundles has no resemblance with that of expert writers and that there is variation between the two disciplines. Another study along the same line is Hyland (2008), who compared uses of 4-word bundles in MA dissertations, PhD thesis and published papers in four disciplines. Despite the proposal of slightly different cluster function categories, the author reached similar results, thus supporting Cortes' findings and reinforcing the role of disciplinary variation.

Several studies have been devoted to the identification of characteristic language patterns that contribute to genres description. Since Swales' seminal book *Genre Analysis* (1990), which focused on the description of research article structure, researchers have been examining this genre and others in order to identify linguistic features that could help to describe such a genre, thus contributing to academic English teaching.

Among innumerable investigations,[45] increasing attention has been given to research grant proposals. For instance, Flowerdew (2016) shares her experience with teaching a research grant writing module for postgraduate students, presenting design and implementation of the classes, drawing special attention to the adoption of a corpus-based approach for carrying out lexicogrammatical analysis tasks. Although this article mostly sheds light on pedagogical activities for raising genre-awareness and formulaic language use, it should be noted one interesting "incidental" finding from the experience. The author reports having unexpectedly encountered instances of non-canonical use of English in the exemplary data source. Especially relevant to my discussion in this thesis are the implications derived from the assumption[46] that speakers

[45] Swales recognises the enormous reach of his work: " the substantial increase in genre studies now makes it impossible in any reasonably sized work to offer a comprehensive review of publications dealing with the academic world in its multifaceted entirety" (Swales, 2004, p. 2).

[46] Flowerdew (2016) used three sources for her course: 1. segments of European Union proposals, taken from Connor & Mauranen's (1999) paper, which were mostly written by Finnish scientists, meaning that nationalities of other proponents are unknown; 2. articles from the Michigan Corpus of Upper-level students paper (MICUSP) corpus, comprising native and non-native speakers of English; and 3. the Corpus of Research Articles (CRA), with articles from 39 disciplines published in prestigious journals and no authors identification. Irrespective of the controversial assumption that nationality indicates someone's mother tongue (the case of this author, who is a Brazilian-Italian national, illustrates the

of English as a lingua franca were the ones producing instances of non-standard English. Such a somewhat biased remark raises some thought-provoking questions: from the standpoint of corpus linguistics principles, does it make sense to attempt to explain the occurrence of such a phenomenon via someone's mother tongue when the purpose of the corpus analysis was the identification of recurrent patterns of language use in those genres in the first place? In other words, is it theoretically viable to "separate the wheat from the chaff"? Matters of corpus linguistics principles will be dealt with in the next chapter (3) and the discussion concerning the role of corpora in this thesis will be presented in Chapter 4.

Continuing in the same methodological vein as Flowerdew's study, Lee and Swales (2006) gave an account of the experience of adopting a corpus analysis approach in classes for non-native speaker doctoral students in an American university. Charles (2016), on the other hand, taught doctoral students (irrespective of their mother tongue) from an English university how to use corpus tools to edit their own texts. In nine courses, a total of 66 students attended her classes and affirmed finding this method useful for their academic writing activities.

Some of the studies presented so far examine the interaction between genres and disciplines, disciplines and language proficiency, and disciplines and writing expertise. Staples, Egbert, Biber, and Gray (2016) instead undertook research in which the development of university students' writing is examined as they advance their studies, ranging from first-year text productions to graduate level, while also taking into consideration the roles played by disciplines and genres in such language description. The results confirmed the authors' hypothesis that advanced writers make more use of phrasal structures, while novices tend to employ clausal structures. Nevertheless, it should be noted that discipline and genre variations also play an important role, as has been shown, for instance, in texts produced in arts and humanities that make more use of clausal constructions than in life sciences and physical sciences.

point), it seems fair to question whether it would be possible that speakers of English as a mother tongue have produced such instances. Although this is not the space for such discussion (which would unveil issues regarding the status of "nativeness"), I claim that it merits further reflection from corpus researchers since that assumption seems to suggest a corpus analysis perspective based on *aprioristic* "truths".

Notwithstanding having received much less attention than its counterpart, spoken academic language has also been investigated. With regards to an examination of lexical bundles, the full-length book by Biber (2006), and the articles by Biber and Barbieri (2007) and Biber et al. (2004) , which sought to study differences and similarities between university spoken and written registers, are references in the area. Another research question concerns vocabulary coverage (see next paragraph), which is a still underexplored area in spoken academic language. Dang and Webb (2014) set out to explore the vocabulary size that is needed by speakers of English as a foreign language to be able to follow lectures and classes, also evaluating to what extent Coxhead's (2001) Academic Word List (AWL) contributes to that. The findings have indicated that with the mastering of this list, students would need to know some 3,000-word families plus proper nouns and marginal words to reach 95% coverage of (oral) texts, while without the list, the number of words would increase to 4,000.

This previous study concerns a highly-valued issue, especially within the pedagogically-oriented research community, like English for Academic Purposes (EAP) and English for Specific Purposes (ESP): academic vocabulary. Among many topics of interest within this scope, it should be noted the weight given to vocabulary learning (teaching strategies, learners' vocabulary knowledge, vocabulary coverage and text comprehension, etc.), as can be seen in Nation's (2001) authoritative book *Learning Vocabulary in Another Language*, whose chapter 6 is devoted to academic vocabulary. One of the most popular approaches for fostering vocabulary learning are word lists, as the proposal of an Academic Collocation List by Ackermann and Chen (2013), clearly demonstrates, focusing on multi-word sequences, and compilation of theme-oriented lists of single units, like the word lists of agriculture (Martínez, Beck, & Panza, 2009) and chemistry (Valipouri & Nassaji, 2013) research articles.

Nevertheless, there has been a dispute relating to the definition of academic vocabulary and the validity of the methodologies of compilation adopted. Hyland and Tse (2009), for example, questioned the existence of an academic core vocabulary. The authors analysed the items on the Academic Word List (AWL, Coxhead 2000) across a number of texts from different areas. Their findings have revealed compelling arguments for reassessing such a taken-for-granted belief:

> corpus analysis shows that individual lexical items on the list often occur and behave in different ways across disciplines and that words commonly contribute to 'lexical bundles' which also reflect disciplinary preferences. (Hyland & Tse, 2009, p. 111)

Many other studies have been devoted to the creation of academic vocabulary word lists and the examination of their controversial efficiency, as will be shown in subsection 2.4 below.

This subsection has given an indication of common objects of study in academic English. Furthermore, it has been shown that academic language is not monolithic, i.e. it varies widely across genres, disciplines, writing expertise and language proficiency; it can be examined from diverse theoretical perspectives; and it is characterised by certain recurrent language patterns.

The next subsection is devoted to what is here called academic Portuguese and the presentation of some research that has been carried out in Brazil and Portugal.

2.3 Academic Portuguese

As shown above, Halliday and Martin draw attention to the fact that scientific language has distinctive features. However, it is still a kind of English. That is why it is a language variety. The authors go on to affirm that the same reasoning applies to other languages, hence "scientific Chinese is a kind of Chinese" (Halliday & Martin, 1996, p. 4). This understanding, by the same token, confirms the validity of the claim for such a register of language and informs that each language (English, Chinese, Portuguese, etc.) has its own scientific language, with likewise distinctive features.

In the same vein, Swales not only acknowledges academic language as a variety occurring in several natural languages, but also argues for "research and development programs for academic languages other than English" (1997, p. 379).

Taken together, these previous assertions strengthen my claim for the establishment of the term *academic Portuguese* as a shorthand to refer to the register of the Portuguese language constructing academic texts (within the epistemological viewpoint presented in 2 above), which in Brazilian and Portuguese higher education institutions are usually used for, among other purposes, acquiring, imparting, proving,

evaluating, discussing, questioning, using, and assessing learning of scientific knowledge.

Moreover, I will show next that academic Portuguese, as understood in this thesis, has been the focus of study of many investigators. Notwithstanding recurrent references to academic Portuguese as still quite incipient in Brazil and Portugal (cf. Carvalho, 2013; Cristovão et. al., 2015; Molsing & Perna, 2014) (especially when compared to academic English), numerous and interesting research on this register is available which demonstrates, on the one hand, acknowledgement of the importance of academic Portuguese, and on the other, the beginning of an effort to characterization.

In the next subsections, I summarise some studies that contribute to the characterization of academic Portuguese, providing longer descriptions of research that I consider elucidative of such characteristics. In addition, I review current projects fostering academic Portuguese teaching and learning and/or developing publicly accessible[47] auxiliary pedagogical materials.

2.4 Academic Portuguese studies

According to Molsing and Perna, "In the broad area of linguistics, Portuguese, in both European and Brazilian varieties, has become one of the most-studied "understudied" languages" (2014, p.1). This observation also concerns academic Portuguese, since research has noticeably increased in both Brazil and Portugal, as the following review shows.

It should be noted, however, that the greatest surge is seen in Brazil, which is not surprising, given the difference between Brazilian and Portuguese higher education scenarios. As demonstrated in Chapter 1, Brazil is facing massification of access to tertiary education at the moment, while Portugal had to deal with that issue some 30 years ago. Furthermore, the overwhelming size of the university student population in

[47] As teachers, we are aware of the production of great teaching and learning resources by those administering courses. However, these are usually individually-applied materials. At most, there is a sharing and exchange at departmental level. Nevertheless, experience shows that such a habit is not frequent. This situation is especially true for languages like Portuguese, with a lack of existing pedagogical materials.

Brazil– nearly 8 million[48] enrolled students in 2016 – indicates a pressing need to address academic literacies in order to promote qualified education to these new students, posing a challenge to educators.

At this point, it should be stressed that there is a significant difference between academic Portuguese studies in Brazil and Portugal, deriving from different modes of research. One way of looking at them is through Bhatia's (2004) categorization of different frameworks used for discourse analysis, in which discourse is understood as "refer[ring] to any instance of the use of written language to communicate meaning in a particular context" (op. cit., p.18), and identified as *discourse as text*, *discourse as genre*, *discourse as professional practice*, and *discourse as social practice* (p. 18). Table 2.1 explains the focus of the four viewpoints.

Table 2.1 Bhatia's categorization of discourse analysis frameworks and focuses (Source: Bhatia, 2004, p. 3)

Framework	Focus
DISCOURSE AS TEXT	lexicogrammatical and other textual properties
DISCOURSE AS GENRE	regularities of organization of language use
DISCOURSE AS PROFESSIONAL PRACTICE	situated language use in institutional, professional or organizational contexts
DISCOURSE AS SOCIAL PRACTICES	language use in a variety of broadly configured social contexts, often highlighting social relations and identities, power asymmetry and social struggle

A literature review has shown that while studies in Brazil address academic Portuguese from all four points of view, in Portugal they clearly favour the discourse as genre, discourse as professional practice, and discourse as social practices frameworks.

For my thesis, it could be considered more relevant to review previous contributions that followed a discourse as text perspective[49]. Thus, this will be the major focus of the research presented next. However, as this kind of investigation is disfavoured in Portugal, other studies concerning academic Portuguese will be reviewed. After all, they not only indicate interest in the topic but also expose an important gap.

[48] As a measure of comparison, Portugal has 10 million inhabitants.

[49] It should be noted, however, that as stated in 2 above, my understanding of academic language implies an interrelation between the four spaces proposed by Bhatia. Nevertheless, since the purpose of academic Portuguese analysis in this thesis is for dictionary-making, then only a stretch of language is described: that encompassing the lexicogrammatical level.

2.4.1 Academic Portuguese studies in Brazil

Within the project called Initiatives of reading and writing in Higher Education in Latin America (ILEES)[50], Cristovão et al. (2015) undertook the identification of Brazilian research groups on literacy in Portuguese (as a mother tongue) in higher education. This first stage of the project adopted a mixed methodology comprising personal interviews with three renowned Brazilian researchers (whose identities are not revealed), search for keywords on the CNPq[51] Working Groups database[52], and universities' web page inquiries.

Overall, their findings indicated the existence of research groups on tertiary education literacy distributed among federal and state universities, covering four out of the five geographical areas of Brazil (the North had no representation). These groups declared their focus to be on either academic literacy, school literacy, teacher literacy, literacy, or on multiliteracy (Cristovão et al. , 2015, pp. 94–96).

In addition, Cristovão and colleagues used the answers provided by the interviewees on referential researchers in the area of academic writing to propose theoretical frameworks currently guiding studies in Brazil, which are summarised in Table 2.2.

Table 2.2 Referential researchers cited by interviewees and the corresponding theoretical framework (Source: Cristovão et al, 2015, p. 89)

Names cited by the interviewees	Corresponding Theoretical framework
Motta-Roth (2008)	Genre Critical Analysis
Kleiman (2005); Koch (2006)	Textual Linguistics
Bazerman (1988)	Genre Rhetorical Studies
Bronckary (1999); Schneuwly and Dolz (2004)	Sociodiscursive Interactionism
Martin (1985, 1989)	Systemic Functional Linguistics
Ferreiro (1981)	Psycholinguistics
Swales (1990)	English for Specific Purposes

[50] http://portuguese.ilees.org/

[51] CNPq stands for *Conselho Nacional de Desenvolvimento Científico e Tecnológico* ('National Council for Scientific and Technological Development'). It is an agency of the Brazilian Ministry of Science, Technology, Innovations and Communications, whose main objective is to foster scientific and technological research and training of Brazilian researchers. For more information, see: http://www.cnpq.br/

[52] According to the authors, this is the website used: http://dgp.cnpq.br/dgp/faces/consulta/consulta%20parametrizada.jsf

Cristovão and Vieira (2016) went on to report on the results of the second stage of the ILLES project. Besides presenting qualitative analyses of the three interviews already mentioned in the previous paper (as mentioned above), their findings indicate the existence of five writing centres in Brazilian universities. A table with the identification of the universities, names of centres and coordinators is reproduced from the original paper here (Figure 2-1):

Universidade	Centros de escrita	Líderes
Universidade Federal de Santa Maria – UFSM	LabLeR – Laboratório de Pesquisa e Ensino de Leitura e Redação	Désirée Motta-Roth e Graciela Hendges
Universidade de São Paulo – USP	Laboratório de Letramento Acadêmico	Marília Mendes Ferreira e Eliane Gouvêa Lousada
Universidade Estadual de Ponta Grossa – UEPG	Laboratório de Estudos do Texto	Djane Antonucci Correa
Universidade Fedral de Santa Catarina – UFSC	Cátedra Unesco	Emérita Leonor Scliar-Cabral
Universidade Federal da Paraíba – UFPB	Cátedra Unesco	Regina Celi Mendes Pereira

Fonte: ILEES Brasil, 2015 (http://portuguese.ilees.org/)

Figure 2-1 Reproduction of Quadro 1: Relação de Universidades, Centros de Escrita e seus líderes (Cristovão & Vieira, 2016, p. 214)

Work developed at some of these centres and within the ILEES project can be seen in the thematic issue of the prestigious academic journal *Ilha do Desterro* on Higher Education Studies in Latin America.[53] Edited by Charles Bazerman, who is the coordinator of project ILEES, and Maria Ester W. Moritz, this special issue "documents, exhibits, and carries forward the desire to understand and improve academic writing at the university level" (Bazerman & Moritz, 2016, p. 9). Contributions cover varied topics, with the final three articles giving an account of some of the outcomes of project ILEES.

Among these centres, I briefly present here the seminal Reading and Writing Research and Teaching Laboratory (REWRITE - *LabLeR*) at the *Universidade Federal de Santa Maria* (UFSM), which was the first of its kind in Brazil, long anticipating the concern with the role of Portuguese in tertiary education. Professor Désirée Motta-Roth, who is a renowned Brazilian expert on academic literacy, is responsible for the

[53] https://periodicos.ufsc.br/index.php/desterro/issue/view/2360/showToc

foundation and co-coordination of this centre which develops research on academic literacy and has been offering academic writing courses to undergraduate and graduate students since 1994 (Motta-Roth, 2012). One key contribution deriving from many years of research and teaching at REWRITE is the book *Produção Textual na Universidade* (Motta-Roth & Hendges, 2010),[54] which is a manual for academic writing in Portuguese inspired by Swales and Feak's book "Academic Writing for Graduate Students".

At this point, it is important to stress that, although Cristovão and colleagues identified ten research groups and five writing centres in Brazil, the existence of other centres and research groups investigating academic Portuguese from diverse theoretical viewpoints and producing research-informed teaching materials is unquestionable. I review some more of them below.

The *Portal da Escrita Científica USP-São Carlos*[55] ('Portal of Scientific Writing USP-São Carlos') is organised by the academic community of the Campus of the University of São Paulo (USP) in the city of São Carlos. It hosts an impressive number of materials (e.g., (video) lectures, courses), bibliographies, and resources (e.g. computational tools, writing tutorials) developed by academics working at that campus. Aiming to support academic writing in English and in Portuguese, there are also links to external resources of interest to writers of academic texts, like rules of text formatting and online lexical resources.[56] Among a series of extremely useful materials for academic Portuguese, I draw attention here to three solid theoretically-grounded resources representative of the educational purpose of the portal.

The first one is the auxiliary academic writing tool Sci-Po, derived from Valéria D. Feltrim's PhD thesis,[57] which helps students of Computational Sciences

[54] It should be pointed out that at times, presentation of linguistic information (for instance, a table containing frequency of occurrences and kinds of quotation verbs per discipline, p. 99) is not based on an analysis of academic Portuguese texts, but rather, a translation of research findings on academic English. CoPEP could provide authentic descriptions of linguistic features that are currently only available for English so that they are included as extra information in DOPU.

[55] http://www.escritacientifica.sc.usp.br/

[56] Interestingly, no links to online dictionaries of Portuguese are available.

[57] As mentioned on the Sci-Po website, the PhD thesis "*Suporte Computacional à Escrita Científica em Português*" was developed at NILC (ICMC - USP/São Carlos), under the supervision of Prof. Maria das Graças Volpe Nunes and Prof. Sandra Maria Aluísio.

compose abstracts and introductions in Portuguese and is freely available online.[58] Besides guiding the writer with the structuring of texts, Sci-Po provides sentences extracted from academic texts, thus showing real examples of language use.

The two other materials co-authored by Jane Raquel Silva de Oliveira and Salete Linhares Queiroz were developed within the scope of a PhD thesis on teaching writing in chemistry. Students (and teachers looking for help with class preparation) find two pedagogical manuals on rhetorical aspects of scientific texts[59] and the structure of laboratory reports.[60] One of the highlights is the use of authentic excerpts from Brazilian academic journals of chemistry to demonstrate real language use.

At the *Pontifícia Universidade Católica do Rio Grande do Sul* (PUCRS), the Research Group on the Use and Processing of Additional Languages (UPLA),[61] coordinated by Professor Cristina Becker Lopes Perna, aims at establishing a tradition of linguistic investigations focused on Brazilian Portuguese for academic purposes, based on a specially compiled corpus of academic textual genres. Molsing and Perna (2014, p.4) report on the focus of interest in some of this group's previous studies: the pronominal system (Molsing & Perna 2013) and adjectives (master's thesis by Beatriz Moro, 2014). The authors also presented ongoing studies at the time of writing (2014), which have since been finished: noun phrases (PhD thesis by Sheila Nunes Santos, 2015), hedges and hedging strategies in L3 Portuguese (master's thesis (2011) and doctoral study by Sun Yuqi, 2015), and metaphors (master's study by Leticia Presotto, 2016).

Growing interest in discussions on teaching and learning Portuguese as an additional language (PAL) at PUCRS, and the need for a specialised space that brings together researchers to share results and classroom experiences have led the *Brazilian English Language Teaching* + (*BELT*+) *Journal* to expand their scope, thus publishing one issue per year on PAL. In the inaugural edition, the editors called for recognition of academic Portuguese as an object of study in its own right. They announced that their purpose was to:

[58] http://www.escritacientifica.sc.usp.br/scipo/

[59] http://www.gpeqsc.com.br/sobre/manuais/jane/Manual-Retorica-do-Texto-Cientifico.pdf

[60] http://www.gpeqsc.com.br/sobre/manuais/jane/Manual-Relatorio-de-Laboratorio.pdf

[61] For details in Portuguese, see http://dgp.cnpq.br/dgp/espelhogrupo/0705653235733182.

> discuss the founding of Portuguese for Academic Purposes as a new area of research with repercussions for how we approach Portuguese for foreigners in academic contexts on the institutional, pedagogical as well as linguistic levels (Molsing & Perna, 2014, p. 1)

In the same fashion, but focussing on another aspect of academic language studies, that of the examination of language learners' errors, Lianet Sepúlveda Torres, Roana Rodrigues and Sandra Maria Aluísio compiled a learners' corpus, called the *Espanhol-Acadêmico-Br* corpus, composed of introductions of academic texts written in Portuguese by Spanish-speaking graduate students[62] at the *Universidade de São Paulo* (USP) in São Carlos. The main objective of this still growing corpus is to "formaliz[e] a typology of the main errors that Spanish speakers enrolled in Brazilian graduate programs make when writing their theses and dissertations in Portuguese" (Torres, Rodrigues, & Aluísio, 2014, p. 100).

As seen so far, there are many perspectives and focuses of research presently being undertaken in Brazil. Another approach that significantly contributes to the description of academic language and the provision of teaching and learning material is Systemic Functional Linguistics (SFL).

Specifically, Motta-Roth and Barbara (2012)[63] draw attention to Brazil's participation in the international project Systemics Across Languages (SAL), coordinated by researchers from the Polytechnic University of Hong Kong, Christian Matthiessen and Kazuhiro Teruya, and Leila Barbara, from the Catholic University of São Paulo. The Brazilian branch (SAL-Brazil), composed of renown linguists from an additional eleven Brazilian universities, "aims at producing an in-depth description of written Portuguese in different genres, using corpus linguistics as a methodological procedure. (Motta-Roth & Barbara, 2012)".

Among a series of PhD theses, MA dissertations, papers and conference communications stemming from this project, I highlight here a few works in order to

[62] At that time, they collected textual productions of students enrolled in the School of Engineering and the Institutes of Physics, Chemistry, Mathematics and Computer Science and Architecture, and Urbanism at USP in São Carlos. However, they intend to extend the compilation of the corpus. (Torres et al., 2014, p.100).

[63] In the foreword of the especial edition of the Brazilian academic journal DELTA on SAL. Available at: http://www.scielo.br/scielo.php?script=sci_issuetoc&pid=0102-445020120003&lng=en&nrm=iso

provide a bird's-eye view of some of the areas of interest investigated within the project so far. All studies make use of corpora comprising academic texts.

Considerable studies have been devoted to the area of letters. For instance, Ninin and Barbara (2013) focused on undergraduate students, examining how they present theoretical voices in their final B.A. thesis by analysing the most frequent lexicogrammatical resources used for that purpose, whereas Ninin (2014) undertook an analysis of 80 theses and dissertations in order to verify the use and functions of the modal verb *poder*. Turning to differences in writing expertise, Ninin, Joseph and Maciel (2015) identified and investigated the use of grammatical metaphor in final B.A. thesis, MA dissertations, PhD theses and academic papers as input for suggesting writing activities. With the same teaching-oriented purpose, Ninin (2015) presented a full class plan with well-designed activities for fostering academic writing at graduate level, solidly grounding her suggestion on SFL principles.

The study of the clitic *se* used for impersonalization in scientific papers from varied areas of knowledge is the topic of Morais' PhD thesis (2013) and follow-up articles on the same subject (Morais, 2016a, 2016b). In addition, the researcher has studied the types of sayers, participant agents of sayer verbs (according to SFL theory) (Morais, 2014) and the function of the verb *haver* (Morais, 2015).

Barbara and Macêdo (2011) set out to study realization patterns of verbal processes in scientific articles written in Brazilian Portuguese comprising a corpus of 1,125 texts (5,176,335 tokens)[64] covering varied areas of knowledge. The verbs *sugerir*, *afirmar*, *propor*, *explicar*, and *discutir* (here presented in order of occurrence in the corpus) were analysed because they are considered, in SFL theory, as the verbs that actualise verbal processes, used for referencing previous and someone else's work, thus typical of scientific texts. The authors examined the forms the verbs most often take and the messages they convey, which can be seen in Table 2.3.

[64] Extracted from journals from different areas on the platform SciELO Brazil. As this platform is the source of texts for CoPEP, detailed information is provided in Chapter 5.

Table 2.3 Verbal processes: top five verbs, verb forms taken and pattern of message realization (Source: Barbara&Macêdo, 2011)

Verbs	Top frequent verb form	Pattern of message realization
Sugerir	3rd person, present	Report - *Que* + indicative^subjunctive and Nominalization
Afirmar	3rd person, present	Report - *Que* + indicative^subjunctive
Propor	Past participle	Nominalization
Explicar	Infinitive	Nominalization
Discutir	Past participle	Nominalization

The authors declared that this study yielded surprising results, namely, verification of changes in established uses or meanings of two verbs: *afirmar* and *propor*. Notwithstanding the understanding of *afirmar* as declarative with performative meaning - thus avoided in scientific texts, according to SFL - this verb ranked second in the frequency list of verbs actualizing verbal process. The other verb, *propor*, appears to be used interchangeably with *sugerir*, which was also an unexpected use, with differences lying in each verb's constituents and pattern of message realization.

These findings are evidence of the need to study and describe academic Portuguese as a variety with its own characteristics. While in academic English the equivalent of *afirmar* has a certain meaning that does not fit with scientific texts, in academic Portuguese not only does it have both a different meaning and frequency of occurrence, but also appears to be a representative candidate for this variety of language. This conclusion should be echoed in instructors' attitudes and the design of teaching-material, as it is clear that the application of distinctive features of academic English to Portuguese writing courses might be quite misleading.

Another source of highly informative descriptions of academic Portuguese is the research project TEXTECC[65] – *Textos Técnicos e Científicos* ('Technical and Scientific Texts'), which encompasses, among others, the subprojects TEXQUIM (about Chemistry) and TEXPED (about Paediatrics) and is carried out at *Universidade Federal do Rio Grande do Sul* (UFRGS) under the coordination of Professor Maria José Bocorny Finatto. As indicated on their website,[66] this enterprise focuses on technical and scientific languages - without restriction to terminology - through analysis of

[65] http://www.ufrgs.br/textecc/

[66] http://www.ufrgs.br/textecc/textquim/

academic texts like papers and textbooks. The main purpose is the development of online tools and dictionaries (the existing ones display meaning and excerpts of authentic texts in two languages)[67] to help undergraduate students of translation studies. Small thematic-built corpora, accompanied by online corpus query tools, are available as well as writing tutorials and lexical resources.

Among numerous interesting works published within this project, I draw attention here to some studies that specifically corroborate the need for a dictionary such as DOPU, namely, those which demonstrate specifications of meanings and uses of linguistic resources due to the scientific character of the texts analysed.

Ramos and Finatto (2012) present a comparison of verbal patterns in texts about paediatrics against patterns described in *Houaiss* - Electronic Dictionary of Portuguese (a dictionary of reference in Brazil). Results revealed a lack of occurrence in the analysed texts of the majority of the senses presented in the dictionary, whith the presence in the texts of some senses not described in *Houaiss*.

Another insightful focus of study is the expression of causality in scientific texts, either via verbs (Alle, 2009; Evers, 2009) or via connectors (Finatto, Evers, & Alle, 2009, 2010). While all investigations share a common objective of helping (novice) translators, the description of the way Portuguese is used in the corpora analysed contributes significantly to the characterization of academic Portuguese. Findings indicate specific employment of verbs and connectors of causality not only diverging from 'ordinary' language use but also presenting further specialization due to genre and disciplinary variation.

For instance, part of the investigation in Finatto et al. (2010) aimed to verify whether some of the connectors defined as expressing causality in previous works (Finatto& Simioni, 2007; Finatto et al., 2006, as cited in Finatto et al., 2010, p. 161) would also convey causality in texts from different areas of knowledge (Chemistry, Physics and Paediatrics). Keyword in context (KWIC) searches of *devido, assim, pois, como, portanto, logo, porque* and *então* yielded 873 lines of concordances. However, careful manual analysis of each concordance by four researchers revealed only 309 occurrences of those connectors were employed with causality senses. Furthermore,

[67] http://www.ufrgs.br/textecc/textped/Dicionarios/DicPed/Principal.php

theier findings suggested that area of knowledge influences the use of connectors: *pois* was the top connector in physics and paediatrics; *devido* was the most frequent word in chemistry.

A contrastive corpus-based study carried out by Finatto and Huang (2005) on the use of adjectives in medicine and chemistry texts is of particular interest due to the proposition of classification of adjectives into three types - discursive, terminological and semi-terminological - and detailed analysis of both collocational behaviour and the morphological constitution of adjectives.

Besides the observation of discrepancy in the frequency of use between the two areas of knowledge, with medicine texts presenting five times more adjectives than chemistry texts, quantitative and qualitative analyses demonstrated variation between the areas in types of adjectives used, recurrent patterns, and morphological constructions.

The classification of adjectives into three types was built on previous work (Finatto, Huang & Enzweiler, 2003, as cited in Finatto & Huang, 2005, p.46) incorporating the basic semantic division proposed by Estopà (2000) in *qualifying* and *relational* adjectives. Thus:

> Type 1) discursive adjective – defined in previous work as "used in daily, common language". Now it is associated with *qualifying* adjective, *i.e.* one which qualifies, describes or characterises the noun. It can be pre- or post-nominal and graded.
>
> Type 2) terminological adjective – earlier associated with a specific area of knowledge. Now it pertains to the *relational* adjective category only, that is, it sub-classifies the noun. It is seen as discipline-determined.
>
> Type 3) semi-terminological adjective – previously defined as adjectives from ordinary language which gained specialized meanings due to employment in scientific contexts. Now it is considered a "variable" adjective because it might be used as a discursive or terminological adjective, depending on the context in which it is employed.

The following Table 2.4 sums up the findings.

Table 2.4 Results of Finatto and Huang's (2005) study on use of adjectives in medicine and chemistry texts (Source: Finatto and Huang, 2005)

Questions	Results
1. Adjectives: classification	*Medicine: mostly use of discursive adjectives. Example: *Assim, por exemplo, a produção de anticorpos e a proliferação de células T tornam-se mais **eficientes** em temperaturas corporais mais **altas** do que o normal.* *Chemistry: mostly use of semi-terminological adjectives. Example: *A maioria das reações são feitas numa ampola a volume **constante*** Terminological adjectives: used relatively little
2. Recurrent associations	In terms of number of adjectives: *Medicine: many cases of noun+adjective+adjective. Frequently, associations of different types of adjectives. Example: *imunidade **celular ativa*** *Chemistry: majority of noun+adjective associations In terms of recurrence: *Some types of adjectives tend to co-occur more frequently with the same adjectives and nouns. Example: *resposta imune adaptativa*
3. Adjectives: morphological construction	*Medicine: more polymorphic adjectives, mostly formed by the process of juxtaposition or agglutination of two lexical bases. Example: *cito+tóxico*

The findings from the studies developed within the TEXTECC project clearly reiterate my understanding of specific use (e.g. collocational behaviour, syntactic patterns) and specific meanings (e.g. verbs, connectors, adjectives) of lexical units in academic texts. In addition to contributing to the characterization of academic Portuguese, these outcomes also reinforce the non-monolithic character of such a register, since genre and disciplinary variations have a strong influence on uses and meanings of linguistics resources. Moreover, the outcomes presented above can be of great help to the process of DOPU's entry-writing as secondary sources.

At the same university (UFRGS), the TERMISUL group, coordinated by Professor Cleci Regina Bevilacqua, has been contributing to the terminology of Portuguese and other languages for over 25 years. Although mostly concerned with terms outside the scope of this thesis, the group has created projects to foster academic writing in Portuguese and in other languages. For instance, while the Learning Object called *Estudo da linguagem científica para o acadêmico de Letras* ('Study of scientific language for humanities students') aims at providing corpus-based activities for raising diverse genre awareness through structure and lexicogrammatical analysis, the project *Estudo de padrões linguísticos de resumos de artigos científicos* ('Study of linguistic patterns in research articles abstracts'), as the name shows, focuses on the abstract genre.

Within the scope of this previous project on the abstract genre, especially relevant for my research is the investigation undertaken by Kilian and Loguercio (2015) about the use of phraseology in research article abstracts written in Portuguese across three disciplines, namely, linguistics, materials engineering, and economics. The object of study was narrowed to what the authors called *genre phraseologies*, understood as non-terminological phraseologies that contribute to characterizing the specific genres under analysis. In this sense, these phraseologies can be considered typical of scientific language. Nevertheless, they may vary according to disciplinary specification.

Adopting a mixed methodology consisting of manual and computational text analyses, the authors first identified informative functions structuring the genre (e.g., to present objectives; to describe methodology, etc.), next examined sequences of 3 to 6 words, utilizing word-specific and n-grams searches, then classified the identified phraseologies according to their functions in the texts.

Overall, the results confirmed that while certain structures typical of the abstract genre transverse all three disciplines, others are employed only in certain disciplines. Discipline specification was also marked by the functions of such bundles, as not all three had the same genre macrostructure.

Table 2.5 below summarises the main findings and provides some examples. In order to facilitate the presentation, I added a *status* category as a shorthand for conclusions provided as full sentences. When no examples were given by the authors, I have opted to leave the cell empty:[68]

[68] Although examples had been given throughout the text, they were not presented in the conclusions as a means to support the point of function exclusivity. Since this particular part appears to be somewhat confusing, I opted to reproduce only what has been stated.

Table 2.5 Conclusions of Kilian and Longuercio's (2015) study. (Source: Kilian & Longuercio, 2015, pp.263-264)

Phraseology	Status	Discipline	Informative Function
o objetivo deste trabalho é	Transdisciplinary	Linguistics Economics Materials Engineering	Objectives presentation
do ponto de vista *sob uma perspectiva*			Theoretical framework presentation
os resultados mostraram que (past tense)	Disciplinary variation	Materials Engineering	Results presentation
os resultados mostram/mostraram (present and past tenses)		Linguistics Economics	
para a análise dos dados	Disciplinary exclusivity	Linguistics	Methodological description
neste trabalho foi/foram feito		Materials Engineering	
os melhores resultados obtidos foram (passive voice)			Results presentation
de acordo com os resultados		Economics	
----------	Function exclusivity	Linguistics	Theoretical framework presentation[69]
----------		Economics Linguistics	Hypotheses presentation
----------		Materials Engineering	Justification[70]

One of the lessons taken from this study is that phraseological analyses cannot rely on frequency statistics alone, requiring also context scrutiny, since high-frequency words have proven to be polysemous. For instance, it was verified that in economics the word *trabalho* ('work') refers to *artigo* ('article') or *estudo* ('study'), but also to the concept *trabalho* as *força de trabalho* ('workforce'), thus considered as a term (i.e. terminology). In other words, investigation of genre phraseologies requires context analysis. Another important lesson refers to the reason for the disciplinary variation. As Kilian and Longuercio put it "epistemological differences inherent to scientific practices are reflected in the language of genres of different discursive communities within

[69] However, it is affirmed earlier in the text (Kilian & Longuercio, 2015, p.263), that economics also presents phraseologies with this function. I opted to include in this summary table only the conclusions presented as such by the authors (op.cit., 2015, pp.263-264).

[70] However, Figure 1 of the paper (op.cit., p.255) also shows occurrence of this function in linguistics. As stated in the previous footnote, I opted to include in this summary table only the conclusions presented as such by the authors (op.cit., 2015, pp.263-264).

academia"[71] (2015, p. 264), adding that further contextual combination analyses would certainly provide more robust support for this assertion.

These conclusions should be present when developing teaching and learning materials, either for classroom uses like the authors' proposal, or for auxiliary pedagogical resources like DOPU.

One more study needs to be mentioned before I move on to an overview of the studies on academic Portuguese in Portugal, as it provides information on academic Portuguese characteristics in comparison to other registers. Prof. Tony Berber Sardinha, who is responsible for introducing corpus linguistics in Brazil, authors this groundbreaking investigation (Sardinha, Teixeira, & Ferreira, 2015) that examines 4-word lexical bundles (see 2.1 above) from the perspective of their distribution, functions and variation in 48 different registers in Brazilian Portuguese.

The research utilised the *Corpus Brasileiro de Variação de Registro* ('Brazilian Corpus of Register Variation') – CBVR, containing over 5 million words, covering 48 registers, with 20 texts in each register. The written part accounts for 72.6% of the corpus, while 27.4% of the total is represented by oral texts.

The methodology of extraction considered three cut-off limits, which were applied to each register: frequency per million words (pmw) (40 times[72] after norming), minimum number of texts (from 3 to 5), and ceiling cut-off of a maximum of 100. Calculations based on these thresholds (for academic articles, a minimum of 3 texts) showed that the academic articles register, containing 92.148 words, has 33 bundles, or 43.4 bundles per million, which represent 2% of the total number of bundles in the whole corpus.

Analysis of the normed frequency of bundles from a point of view of routinization, that is, "the extent to which the lexical combinations found in a register are repeated across different texts" (Sardinha et al., 2015, p.44), academic articles were included in the lowest level, with 2,843 pmw, marked by less conventional language and little use of bundles. The authors stressed that this unexpected classification, in fact,

[71] "*diferenças epistemológicas dos fazeres científicos refletem-se na linguagem dos gêneros das diferentes comunidades discursivas dentro da academia*".

[72] The authors inform that, since subcorpora sizes varied due to different lengths of the texts, in some cases the thresholds of minimum frequency per million and texts per registers were altered.

needs to be interpreted with caution since there are many variables at play. They suggest that the normed frequency number is comparable to those from similar analyses of Spanish and English texts.

The second analysis focused on the function of the bundles, based on the work of Biber et al. (2004) on categorization of stance, discourse organisers and referential bundles (as presented and explained in 2.1). Academic articles present a majority of referential bundles. The next table (Table 2.6) reveals the top four most frequent constructions found, with their frequency per million words and number of texts in which they occurred.

Table 2.6 Bundles and their frequency in the academic articles genre of CBVR. (Source: Sardinha et al., 2015, p. 45)

Bundle	Frequency per million words	Occurrence in texts
Com o objetivo de	76	6
Da década de N	43	3
De acordo com a	109	4
De acordo com o	119	7

One of the most significant consequences of this study for my research is the confirmation that academic Portuguese is a register different from others, supporting my claim in 2.1.2 with solid evidence. In addition, the conclusion that the taxonomy originally created for English is also applicable to Portuguese is extremely positive, as such classification might contribute to writing entries for DOPU.

Although the previous review has only covered part of the work that has been done in Brazil, it seems possible to suggest that academic Portuguese has become an important topic of research from various theoretical frameworks and for different purposes. It should be noted, however, that among all these studies, none concerns the development of a dictionary, despite the exposition of a series of very particular language patterns in academic Portuguese that tend to pose challenges for students.

The next subsection reviews the status of research on academic Portuguese in Portugal.

2.4.2 Academic Portuguese studies in Portugal

As mentioned at the outset, in Portugal research tends to focus more on genre, professional practices and social practices to the detriment of text and its properties. Although the majority of investigations do not address directly what is required for my thesis, it is important to review a few of them in order to demonstrate the relevance of academic writing and reading in Portugal, thus further supporting the significant contribution that DOPU could offer as an auxiliary pedagogical resource in literacy practices.

One elucidative piece of research on the characterization of academic literacy practices is Fischer (2011). Although her ethnographic-based investigation focused on a textile engineering course, the findings seem to echo many voices often heard in higher education settings in Portugal (cf. Rodrigues, 2009)), especially with regards to professors' opinions, and tensions between students and professors' expectations relating to writing production (see Pereira & Graça (2014) for further discussion on concepts of writing at the tertiary level in Portugal).

In personal interviews with the author, professors expressed a common conviction on students' academic literacy skills as problematic and insufficient; for them, students "produce nothing" during the course (Fischer, 2011, p.42).

Contrary to the professors' views, Fischer's examination of all genres with which students had contact throughout the course revealed that students do produce a great deal, since these genres specialize on academic reading and writing practices. The following table (Table 2.7) shows the genres she identified and their purposes:

Table 2.7 Genres used in academic literacy practices at a course of Textile Engineering. (Source: Fischer, 2011, pp.43-44)

Pedagogical purpose	Genres
Organization of learning and teaching processes; Construction and expansion of knowledge; Mediation of content acquisition	Comprehension of: • Courses evaluation grids; • Syllabi; • Assessment instruments (tests, exams, reports); • Notes on tests and exams dates; • Classes summaries in PowerPoint (PPT) format; • Instruction manuals; • Lists of exercises (calculus); • Models of texts structures (agenda, minutes, reports, problems resolution)
Guiding students' reading practices	Comprehension of: • Chapters of textile manuals; • Online textile catalogues; • Advertisement of textile and electronic products; • Geographic maps; • Images of textile products; • Tables with figures; • Commercial newsletters; • Summaries of classes from other universities in PPT format; • Emails with technical instructions; • Online scientific articles.
Regulation of students' learning	Comprehension and production: • Notes (taken during classes); • Answers in tests, exams and lists of exercises; • Lab and project reports.

Fischer's study also revealed the source of tensions between professors and students' expectations regarding writing activities. While professors affirmed that students have poor academic literacy skills, thus failing to produce satisfactory project reports (in this course, this is the most valued learning assessment instrument), students claimed that there was no explicit instruction, leading them to write project reports in the way they knew how. It can be argued that, at the heart of this tension lies academic language knowledge, which is taken for granted in the professors' view, whereas for students, it is one specific area of knowledge to be developed in the course, hence requiring overt teaching (see also Rodrigues, 2009 for a lengthy case study on polytechnic students' writing difficulties).

By exposing the high degree of autonomous learning of academic language expected from students, Fischer's investigation reinforces the need for academic reading and writing auxiliary pedagogical tools, such as DOPU.

In line with the view that not only do students need support with academic language, but also instructors should know the directions to be taken in explicit literacy practices guidance, the Corpus of Portuguese Undergraduates' Texts (CUTe)[73] has been developed since July 2012 at *Escola de Educação de Lisboa* (ESELx). As an error-tagged learner corpus of Portuguese academic writing, it aims to provide teachers with more accurate, directed linguistic orientations based on the types of errors made by students (Cardoso, Magro, Braz, & Nunes, 2014). This corpus is hyperlinked with the website *Scriptorium- Centro de Escrita Académica em Português*.[74]

Scriptorium is a 'Writing centre whose mission is to make available those contents and services that promote development in Portuguese academic writing competence' (my translation).[75] Catarina Magro and Adriana Cardoso are responsible for this promising virtual centre, which hosts ever-growing resources (writing guides, tutorials and exercises); activities (workshops, courses, tutorial support); services (proofreading and teaching training); information about research developed by their team and links to external sources, like the CUTe corpus and online dictionaries of Portuguese.

Additional enlightening contributions to the characterization of academic Portuguese derive from the research group Discourse and Academic Discursive Practices at the Centre of General and Applied Linguistics Studies (CELGA-ILTEC), at the University of Coimbra. Within this group, special attention should be drawn to some researchers that have, by adopting the *Interactionnisme Sociodiscursif (ISD)* framework, set out to examine and describe typical academic genres in higher education in Portugal. Among their studies, some involved investigations on development answers (exam answers) (Silva & Santos, 2011), comparison between abstracts and introductions (Silva

[73] http://www.cute.org.pt/ . As of 20 February 2017, the site informs that CUTe is temporarily inaccessible due to upgrading.

[74] http://scriptoriumescrita.wix.com/scriptorium#!__page-0t

[75] "*Centro de Escrita que tem como missão disponibilizar conteúdos e serviços que promovam o desenvolvimento de competências no domínio da escrita académica em português*"

& Santos, 2015) and between PhD dissertations and research articles (Santos & Silva, 2016).

Another particularly interesting piece of research that supports the view of academic Portuguese as a register of language with its own characteristics is Santos and Silva's (2016) study on PhD dissertations and research articles. One of the stages of the study involved comparison of the results of a manual analysis performed on a corpus comprising 130 dissertations from various areas of knowledge published at the University of Coimbra, using Swales' (1990; 2004, as cited in Santos & Silva, pp.176-180) taxonomy. Among many interesting findings, an especially revealing one stems from the fact that Swales' categorization of PhD dissertations in three structure types diverges from the characteristics found in Portuguese dissertations by Santos and Silva, thus requiring appropriate adjustments. In other words, the very need of taxonomic alteration reinforces the understanding of academic Portuguese as a register with its own rights, which might present common features across languages, but that also has distinctive traits.

From a different theoretical paradigm, Bennett reached a very similar conclusion. Bennett's PhD thesis (2008) approaches Portuguese academic discourse from a contrastive rhetoric perspective, in the area of translation studies. The highly-experienced translator's purpose was to obtain confirmation for her hypothesis of existing differences between academic discourse in Portugal and hegemonic English academic discourse (EAD). Additionally, her two other objectives were related to translation and ideological issues deriving from translation processes. Although the general view of analysis is discourse as social practice (taking on Bhatia's categories as mentioned earlier), my interest lies on the outcomes from the method adopted for proving her hypothesis.

Bennett (2008, also in 2010a, 2011)[76] analysed a corpus comprising 408 texts (1,333,890 words) produced by expert writers of academic Portuguese, in varied genres and disciplines. Each text was analysed and awarded a *Variance Factor* (VF) grade, which reflects the level of distinction between the discourse in those texts and EAD.

[76] Bennett (2010a) and (2010b) are journal articles resulting from her PhD thesis (Bennett, 2008). Despite some variation among them, for instance, in terms of terminology (*e.g.*, Deviation Factor (DF) in her PhD thesis instead of Variance Factor (VF) in her 2010a paper), they all refer to the same study.

Elaboration on VF and discussion of discourse comparison results are beyond the scope of my objective in this present work.

Determination of VF was mainly based on the analysis of *Distinguishing Discourse Features* (DDFs), which are lexicogrammatical aspects of the Portuguese language used in those texts. The corpus analysis that characterize those DDFs are of particular interest to my study and are summarised in Table 2.8.

Table 2.8 Bennett's Distinguishing Discourse Features (Source: Bennett, 2008, 2010a)

Distinguishing Discourse Features (DDFs)	Findings
Personal References (Pers)	Use of personal references: first- and second-person verb forms (singular and plural), personal pronouns, and their respective possessive adjectives; the 'magisterial plural'. Functions: a) Self-reference to author: for signposting; for referring back to points already made; for expressing personal opinions about the matter in hand; b) Impersonal usages: the first-person plural often used in generalised situations; c) Reference to the discourse community: used when the author is identifying with a position that is generally accepted by the discourse community; d) To refer to the here-and-now (deictic use): reference to Portugal using the first-person plural; similar usage as regards time.
Gerunds (Ger)	Used to express a wide range of syntactical relationships, including temporality (anteriority, posteriority, simultaneity), causality, consequence, purpose, condition and concession.
Framing devices (FD)	Assertions and observations tend to be presented indirectly, embedded in a main clause that emphasises the interpersonal dimension. Use of excessively elaborate FDs (e.g., *reveste-se também de particular interesse verificar que*) FDs functions: a) FDs with little or no semantic content: *constata-se que*; b) FDs expressing emphasis: *convém sublinhar que*; c) FDs expressing attitude: *não surpreende que*; d) FDs expressing epistemic modality: *resta-nos a certeza de que*.
Deferred Topic (DT)	The main idea is not placed in initial position, it is deferred. Sentence level: cataphora; Clause level: inversion of the SV word order; Paragraph level: topics deferred to the end of the paragraph.
Complex syntax (CS)	Long sentences with dense subordination. Main clauses interrupted by supplementary information, such as participle phrases, parentheses and lists.
Top-heavy Sentences (TH)	Excessive amount of information between the grammatical subject and the verb.
Verbless sentences (VS)	Usually subordinate clause or participle phrase that has become detached from its main clause.
Multiple Negative Constructions (Neg)	Most common: negative with a lexical item that also has a negative charge.
Historical tenses (HT)	Use of the present or future tenses to refer to events that took place in a contextualised past.
Poetic, figurative or high-flown diction (PD)	Frequent use of high-flown literary style, e.g., the University of Coimbra described as "*instituição mater cujo corpo ilumina o tempo com as luzes do saber*"
Poetic or philosophical quotations (Cit)	Insertion of quotations in the middle of an article.
Abstractions (Abs)	1. Lexical abstractions: addition of Latinate suffixes to existing roots. Particularly frequent noun derivation from adjective basis (e.g., *intencional -> intencionalidade*) 2. Syntactical abstractions: a) collocation of abstract nouns with material and verbal processes (e.g., *a consciência da etnicidade colabora*) b) use of the archetypal singular
Postmodern features (PM)	Use of wordplay, neologisms, paradoxes

Bennett's seminal work is fundamental for my research in many ways. Firstly, it not only supports the understanding of academic Portuguese as a register with its own characteristics but also provides ample evidence to prove it. Secondly, each of the Distinguishing Discourse Features can, in one way or another, contribute to the lexical description of academic language acquired from CoPEP; for instance, the identification of the use of gerunds to express varied syntactical relationships may be revisited when writing an entry. Finally, taken together, these results provide an important insight on what kind of results that might be expected when working on GDEX[77] configurations (more on this in Chapter 8).

This literature review has revealed that a considerable number of studies on academic Portuguese have been undertaken. While the vast majority has been pursued in Brazil, where one can see a wide spectrum of theoretical frameworks and accordingly, objects of interest, research has also been carried out in Portugal, indicating a concern for students' needs.

Furthermore, this section has shown the development of some research-informed teaching material and didactic resources. However, there is definitely a need for more investigation and consequently more publicly available pedagogical material.

It should be noted that, among all the works reviewed, the lack of a dictionary of academic Portuguese stands out, notwithstanding the acknowledgement of its key role in the process of (academic) language learning.

I wrap up Chapter 2 with a brief discussion on academic vocabulary in order to elucidate in what sense DOPU will be a dictionary covering academic language and not (what some consider) academic vocabulary only.

2.5 Academic vocabulary

According to Nation, academic vocabulary is "common to a wide range of academic texts, and not so common in non-academic settings" (Nation, 2001, p. 189).

[77] GDEX stands for Good Dictionary Examples and is a function of the Sketch Engine corpus tool (Kilgarriff et al., 2004). In brief, through a set of heuristics, the system searches the corpus for optimal examples and ranks them, displaying the best ones at the top. Such preselection has proven to streamline lexicographic work to a great deal.

The concern with this issue has resulted in several studies and attempts to produce word lists. Among these, the most well-known work is Coxhead's (2000) Academic Word List (AWL), which has had a major impact in the area of English for Academic Purposes (EAP). Hence, it is not surprising that it has soon become a dominant model for other academic word lists, either with regards to other languages (e.g. Portuguese (Baptista, Costa, Guerra, Zampieri, & Cabral, 2010); Swedish (Jansson, Kokkinakis, Ribeck, & Sköldberg, 2012); Kokkinakis, Sköldberg, Henriksen, Kinn, & Johannessen, 2012); and Turkish, (Dolmacı & Ertaſ, 2016)); discipline specialization (e.g., the Medical Academic Vocabulary List (MAVL), Lei & Liu, 2016; the Environmental Academic Word List (EAWL), Liu & Han, 2015; A Nursing Academic Word List (NAWL), Yang, 2015; Chemistry Academic Word List (CAWL), Valipouri & Nassaji, 2013; Engineering English Word List (EEWL), Hsu, 2014)); or multi-word expressions list (e.g. collocations list, Durrant, 2009; the Academic Collocation List, Ackermann & Chen, 2013; and Academic Formulas List, Simpson-Vlach & Ellis, 2010).

The AWL is composed of 570 word families, that is, "a stem plus all closely related affixed forms" (Coxhead, 2000, p. 218), extracted from the Academic Corpus, which had been specially compiled for this endeavour. The corpus contained 3.5 million tokens and 414 texts from four domain categories (Arts, Commerce, Law and Science), which were further narrowed down into seven subject areas each, amounting to 28 subject areas. Methodology for words extraction was as follows:

> 1. Specialised occurrence: The word families included had to be outside the first 2,000 most frequently occurring words of English, as represented by West's (1953) GSL.[78]
>
> 2. Range: A member of a word family had to occur at least 10 times in each of the four main sections of the corpus and in 15 or more of the 28 subject areas.
>
> 3. Frequency: Members of a word family had to occur at least 100 times in the Academic Corpus (Coxhead, 2000, p. 221).

[78] GSL stands for General Service List and was drawn up by West in 1953 in order to help foreign learners of English.

Despite its extensive application in EAP teaching materials, courses and dictionaries and its undeniable worldwide influence on academic language teaching, the AWL has received sharp criticism.

The first criterion (specialised occurrence) was criticised in at least two ways. Firstly, the general word list employed as exclusion criterion - the GSL - was published in 1953, but based on work from the 1930s (Leech, 2001). That means it contains outdated words, thus not being representative of contemporary language (Gardner & Davies, 2014; Hyland & Tse, 2007). Secondly, the assumption that academic vocabulary should exclude the top 2000 most frequent words of general English because students know them was questioned. On the one hand, Hancioğlu, Neufeld, & Eldridge (2008) and Martínez et al. (2009) claimed that academic texts do contain words from the GSL. On the other, such supposedly known words "may be used differently in academic discourse" (Paquot, 2010, p. 15). However, due to familiarity with them, "learners may not even notice when they have not understood them properly" (Durrant, 2009, p.164). For instance, Durrant (2009) carried out an investigation to verify the collocational use of AWL items and his findings indicated that two high-frequency words from the GSL, 'address' (noun) and 'mean' (verb), have been found in academic texts with other meanings, as the collocation pairs 'address-issue' and 'by-means' clearly demonstrate.

The Academic Corpus used for word list compilation is criticised for having texts taken from old corpora (from the 1970's), and from unbalanced origins (64% from New Zealand, 20% from Great-Britain, 13% from the United States, 2% Canada and 1% from Australia) (Kosem, 2010). Another point that has been raised is that the corpus is biased toward disciplines like business and law (Hyland & Tse, 2007).

Another significant drawback refers to the nature of AWL's construction, namely, word families. Kosem (2010) and Paquot (2010) argued that a headword cluster may include words which are not frequent. In addition to the frequency discrepancy, Ming-Tzu and Nation (2004, p. 294) referred to the case of inclusion of headwords in the AWL, despite the lack of occurrence of some of its family members in the Academic Corpus. The authors also have shown that related word forms under a word

family may not be connected in meaning at all (e.g. *consist* ('contain') and *consistent* ('unchanging')).

The fact that the AWL comprises single items only instead of phraseology (that is, phraseological units, idioms, collocations) has been strongly contested (Hyland & Tse, 2007; Paquot (2007); Hancioğlu et al. (2008); Durrant 2009). Specifically, attention has been drawn to the importance of considering sequences of words due to their salience and functional significance (Simpson-Vlach & Ellis, 2010). Another point that has been made is that different disciplines have different specializations of word meanings and favour different collocations (Hyland & Tse, 2007, 2009). In sum, it can be said that "there is considerable variation in the use and discourse function of phraseology across academic disciplines and registers" (Kosem, 2010, p. 48).

Finally, Paquot (2007, 2010) questioned the value of the AWL for productive purposes. She then compiled another academic word list, the Academic Keyword List (AKL), which included the GSL words, but took the comparison with a fiction corpus as the methodological approach to define the keyness criteria. That means that key lemmas (and not word forms, as in Coxhead's AWL) that appeared in the MicroConcord corpus but not in the fiction reference corpus were selected to be part of her 930-lemma list. Despite trying to overcome some of Coxhead's AWL shortcomings, Paquot's list still shows limitations as to the corpus status (composed of incomplete texts) (Kosem, 2010) and to the fact that her hypothesis of academic words being less represented in the fiction corpus has not been proved (Kosem, 2010, p. 44; Kilgarriff, 2012, pp. 128-129).

Claiming that "there has been much less attention paid to the idea of a dictionary of academic English, as opposed to a word list" (Lea, 2014, p.182), the *Oxford Dictionary of Academic English* (ODAE) (2014) comes out as a fully-fledged dictionary, whose stretch of language covered is said to be defined based on an alternative theoretical perspective. However, it will be shown that this is not quite the case.

As a detailed review of ODAE (see Kuhn, 2015) shows, in its *Introduction* and in *Reference Section 22*, it is explained that academic vocabulary can be divided into

three broad categories: an ordinary general English vocabulary; a specialist, discipline-specific vocabulary; and in between, a general academic vocabulary:

> These are words that tend to be used across most or all academic disciplines; most are also used in general English. However, the way they are used in academic writing is often rather different, which is why these words deserve special study from the student of academic English. It is these 'general academic' words that are the main focus of this dictionary. (ODAE, 2014, p. v, R22).

Firstly, this is the underlying principle foregrounding all the academic lists that have been mentioned so far. It is also clearly based on Nation's categorization of the vocabulary of non-fiction texts as "high frequency (or general service) vocabulary, subtechnical or academic vocabulary, technical vocabulary, and low frequency vocabulary" (Nation, 1990, p. 19, as cited in Nation & Kyongho,1995, p.35), which is grounded in the belief that academic vocabulary should be built on top of the 2000 most frequent words of a language.

As can be seen, by affirming that there is an ordinary general English vocabulary that shall not be included in ODAE, traditional uptake for word list building was still employed. Furthermore, although the earlier explanation suggested a frequency-based approach for nomenclature building, in fact, as I stated elsewhere (Kuhn, 2015, p. 114), it was said that lexicographers set out the headword list compilation based on Coxhead's AWL and Paquot's AKL.

For the Portuguese language, there has been one attempt to compile an academic word list, the Portuguese Academic Word List (P-AWL) (Baptista et al., 2010). This project, however, has serious methodological drawbacks mostly due to the fact that P-AWL is the result of the translation of Coxhead's AWL into Portuguese.

Firstly, AWL has itself suffered severe criticism as stated above. Thus, its reliability as a source list can be questioned. By extension, P-AWL's credibility is also debatable.

Secondly, the translation of decontextualized word forms suggests the understanding of language as transparent, words as non-polysemous, and context (in co-textual, textual, discursive, and cultural levels) as merely ancillary to meaning construal.

This is a highly questionable principle, as has been demonstrated throughout this chapter.

Thirdly, the fact that P-AWL results from translation explicitly indicates that authentic use of academic Portuguese was disregarded. Indeed, the non-adoption of an evidence-based compilation approach goes against the established acknowledgement that "corpora are used, and are now widely accepted as valuable, arguably essential, resources of serious linguistic description of any kind" (Moon, 1998, p. 347). The justification for not using corpora in the project is that "establishing a (balanced) corpus for deriving the vocabulary intersection, across scientific domains and genres, may well be an impossible mission" (Baptista et al., 2010, p. 121). However, it can be argued that "impossible" might be considered a quite strong word, given the several valid academic corpora that have been built, including CoPEP for this thesis. In any event, the fundamental problem lies in the implied reason why the use of corpora was thought to be the preferred option in the first place, namely, due to the need of searching for a cross-disciplinary common vocabulary.

This lack of authentic evidence of Portuguese language use leads to the third limitation the project reveals: subjective adjustment of the translated words by annotators indicate an option for intuition over evidence – a methodological approach that has been questioned since Sinclair inaugurated the tradition of empirical lexical analysis (Hanks, 2008, p. 222).

The uncovering of P-AWL's serious shortcomings raises a question as to its usability for teaching-learning academic Portuguese. As a source for my project, the use of word lists (of any kind) goes against the principles sustaining this PhD research, thus P-AWL will not be used.

Kosem (2010, p. 45) claimed that the major limitation of the corpus-based English word lists that have been compiled so far, and I include here the *Oxford Dictionary of Academic English*, is the exclusion of words (general, high-frequency words; technical words). By following this principle, compilers assume that students have fully mastered these words, when, in fact, he questions whether they would really know everything that knowing a word entails (all meanings, collocations, patterns, synonyms, etc.). That means that:

> It is more likely, however, that students will know core meanings and patterns of the words, *i.e.* meanings which are normally most frequent in general English. Consequently, if a word has a different meaning or different distribution of meanings in academic English, it will probably present difficulties to students. (Kosem, 2010, p. 45)

Todd (2017) shared a similar view on word lists. He understood that polysemous words are opaque and that his students in engineering needed help to learn their meaning in the context of course texts. He thus set out to compare the meanings of words in context against the main meanings given in the online dictionaries that students often relied on in order to prove his point. His findings indicated a discrepancy between the order of the senses presented in the dictionaries and the ideal order to comply with the real frequency of the senses used in their particular course context. Based on that, he proposes a list of opaque words for engineering students. He explains:

> the main criterion for choosing which words and meanings should be included on the final list is opacity. This criterion should identify those words for which the learners would gain the greatest benefit from a teacher's help, since these are the words learners are most likely to have problems dealing with autonomously. The opaque word list consists of fairly high-frequency polysemous words where the meaning required is not the usual meaning associated with that word (Todd, 2017, p.38)

It should be noted that Todd was, in fact, making up for the lack of a dictionary of academic English in which explanations of words are provided according to their use in different disciplines. Such a dictionary[79] would in fact grant students autonomy, letting teachers profit more from class time.

Based on the discussion presented here, the design of DOPU will follow Kosem's (2010, pp. 45-46) approach to lexical analysis for dictionary-making, that is, to conduct a semantic analysis, taking into consideration word context, its associations with other words, among other criteria that are explained in detail in Part III. This methodological approach is in consonance with the literacy-oriented concept of

[79] It is unfortunate that Kosem's proposal of an alternative method for making a dictionary of academic English, allied to a performance of "a semantic analysis of the words, also considering the differences between their roles in general English and academic English" (Kosem, 2010, pp. 45-46), was never put in practice. This dictionary would definitely have been useful for Todd and all other teachers and students around the world.

academic language adopted in my research and expressed throughout this chapter, and is perfectly summarised in this brief excerpt by Hyland and Tse's (2007, pp.236-237):

> It is by no means certain that there is a single literacy which university students need to acquire to participate in academic environments, and we believe that a perspective which seeks to identify and teach such a vocabulary fails to engage with current conceptions of literacy and EAP, ignores important differences in the collocational and semantic behavior of words, and does not correspond with the ways language is actually used in academic writing. It is, in other words, an assumption which could seriously mislead students.

2.6 Summary

The review undertaken in this chapter indicates that there is an academic Portuguese with its own characteristics in terms of lexical and grammatical associations. For English academic language, which has been studied for years, the outcomes of the studies are constantly incorporated in teaching materials and lexicographical resources (*e.g*. *Oxford Academic English Dictionary*, 2015). The same does not apply to Portuguese. Although research has been growing, pedagogical materials are still scarce. Among them stands out the lack of a dictionary for university students. I conclude that a corpus-driven lexicographical work on academic Portuguese is imperative, though yet to be developed.

Chapter 3 Corpora and dictionary-making

This chapter reviews the relationship between corpora and dictionary-making. It starts with an account of corpus linguistics (3.1). A brief history of corpus linguistics is presented in 3.1.1, while in 3.1.2, the controversial discussion on the status of corpus linguistics is addressed. Next, in 3.1.3, different applications of corpus linguistics are presented, and in 3.1.4 corpus characteristics are described, with special attention to corpus design (3.1.4.1) and possible approaches for corpus analysis (3.1.4.2).

Section 3.2 focuses on the electronic revolution and its consequences for corpora and lexicography. It begins (3.2.1) with a succinct history of the making of the Cobuild dictionary and its importance for lexicography. In 3.2.2, an account of the process of corpora enlargement is given, whereas in 3.2.3, the first movements towards greater automation in lexicography are recollected. These advances have led to the development of an area of studies – e-Lexicography -, which is reviewed in section 3.2.4. Finally, in 3.2.5, I introduce a highly innovative approach to dictionary-making, which was employed in this PhD project.

To wrap up this chapter, a succinct review of the history of corpora compilation and Natural Language Processing tools development in the context of the lexicography of the Portuguese language is provided in 3.3, with intermittent highlights of notable dictionaries which were based, to varying extent, on corpora. Attention is drawn to the evaluation of existing corpora of Portuguese containing academic texts in order to verify whether they were suitable for my research.

3.1 Corpus linguistics

In this thesis, corpus linguistics (henceforth CL) is used to provide the guiding principles and techniques defining corpora design, compilation and analysis. It will be interchangeably referred to as ‘approach’ (Biber, Conrad, & Reppen, 1998; Stubbs, 2006) and ‘method(ology)’ (Gries, 2009; Hunston, 2006; Leech, 1992; McEnery & Wilson, 1996; McEnery, Xiao & Tono, 2006; Meyer, 2002). According to Biber et al. (1998, p. 4), the characteristics of a corpus-based approach are:

- it is empirical, analysing the actual patterns of use in natural texts;

- it utilizes a large and principled collection of natural texts, known as a "corpus", as the basis for analysis;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques.

Tognini-Bonelli's (2001, p. 1) characterization of CL is broadly consistent with Biber et al.'s (1998), only with additional stress on the study of meaning:

> It is an *empirical approach* to the description of *language use*; it operates within the framework of a *contextual and functional theory of meaning*; it makes use of the *new technologies*.

Taken together, these characteristics are in consonance with Sinclair's concept of corpus, which is the one followed in this thesis:

> (…) a collection of pieces[80] of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair, 2005).

3.1.1 Brief history of corpus linguistics

Corpus linguistics generally understood as language study based on authentic evidence of language use has a long history. McEnery and Wilson have shown that many researchers had carried out investigations on a variety of subjects such as language acquisition; spelling conventions; language pedagogy; comparative linguistics; and syntax and semantics, utilising corpora of real language use (1996, pp. 2-4) long before the term 'corpus linguistics' began to be actually used, in the mid-1980s (Leech, 1992, p. 105). In fact, Leech (1992, p. 105) suggested that the reason for the lack of such a label was because "for those who espoused this approach, corpus linguistics was simply "linguistics"".

However, with the rise of Chomsky's rationalism-driven theory of language studies in the 1950s and 1960s, a claim was made that competence, not performance,

[80] Sinclair (2005) explains the use of "pieces" due to some researchers who compile extracts of texts (e.g. the Longman/Lancaster English Language Corpus, Summers, 1993).

was the real object of linguistic investigation. Hence, empiricism-based linguistic research, i.e. those using corpora for studying the language, were highly criticised. As McEnery and Wilson (1996, pp. 4-5) explained, according to this new perspective, linguistics should be based on "theories which reflected a psychological reality, cognitively plausible models of language" instead of abstract descriptions of utterances in a corpus, which Chomsky deemed skewed, and thus untrustworthy.[81] With this paradigm shift, Chomsky's generative theory of language took over the position of mainstream scientific linguistics (see Hanks, 2009).

It should be noted that, despite the undoubted predominance of generative linguistics in those decades and the consequent loss of popularity of corpus-based studies, corpus analyses were still being carried out, mostly due to the contribution of technology to the advancement of this kind of research (cf. McEnery & Wilson, 1996; Sardinha, 2000).

In this vein, notwithstanding the theoretically-adverse context of the time, Nelson Francis and Henry Kučera created the 1-million-word Brown corpus (Brown University Standard Corpus of Present-Day American English), covering 15 genres of written American English published in the year 1961, and releasing it in 1964 (Baker, 2011, p.17). This corpus comprised, for the first time, machine-readable texts stored in a computer.

Other corpora compilations with ever-increasing sizes followed suit. For example, the five times larger 5-million-word AHI corpus (American Heritage Intermediate Corpus; written American English, 1971) was an impressive mark in the progress of corpus linguistics; in the United Kingdom, corpora were reaching sizes as large as 20 million words, namely, the Birmingham Corpus (Birmingham University International Language Database; written British English, 1987), used for making the innovative Cobuild Dictionary of English for Advanced Learners from scratch (more details in section 3.2 below).

It is apparent that the analysis of large-sized corpora required other methods than the traditional manual inspection used in hand-picked instances of language use. Fast

[81] See McEnery & Wilson (1996, pp. 4-14) for a detailed account on Chomsky's opposition to corpus linguistics.

developing technology contributed to advances in computer-based corpus analysis tools, for example, concordancers, left and right sorting of keywords in context, search for collocates within a 2 to 5-word span from the keyword, word lists, etc. As can be concluded, the advent of the computer set a clear division mark in the history of corpus-based studies, giving rise to the concept of corpus linguistics as we presently know it.

It was thus possible, for the first time, to undertake a systematic examination of large amounts of evidence of language use that revealed patterns and regularities never before seen (cf. Sinclair, 1991). Quantitative results of language behaviour analyses enabled Sinclair to show that "words are interconnected, not isolates, that meaning is derived from context, and that collocation is key" (Moon, 2008, p.243). With this revelation, Sinclair made a compelling case against principles of generative theory and shed new light on linguistics studies.

One of his most renowned arguments was that meaning construction in texts did not follow the 'slot-and-filler' model proposed by generative theory. That is, he argued that while Chomsky's tradition defined production of meaning through the filling of grammatical slots with virtually any random word, according to speaker's choice, corpora investigation revealed that meaning derives from phrases, which are more or less fixed and have varied extent.

Sinclair's idiom principle, "that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair, 1991, p.110), has thus revolutionised linguistics, opening up unprecedented possibilities for research on how language works.

For instance, it has been shown that phraseology is pervasive and highly frequent in language. Stubbs (2006, p. 24) indicates that Mel'čuk' s (1998) estimates of a frequency of "ten times as many phrasal units as individual words" in language. Gries also shares this view. He succinctly explains how meaning depends on words co-relations:

> (…) formal differences reflect, or correspond to, functional differences. Thus, different frequencies of (co-) occurrences of formal elements […] are assumed to reflect functional regularities, and 'functional' is understood here in a very broad

> sense as anything – be it semantic, discourse-pragmatic, …- that is intended to perform a particular communicative function. (Gries, 2009, p.4)

In this vein, painstaking analysis of lexical items and inspection of recurrent structures in very large amounts of naturally occurring language evidence soon became a method largely adopted in a variety of studies. This is the case of many studies on academic language, as presented in Chapter 2, which use corpora for analysing multi-word units[82] and their meaning, function or specialized use with regard to genre, discipline, writing expertise and language proficiency variation.

As can be seen, corpus linguistics has restored empirical research to a legitimate position within language studies. In consequence, corpora have been exponentially increasing not only in size – mega corpora, as it will be shown later, are reaching sizes of 20 billion words – but also in language coverage. While in the beginning of electronic corpora, the vast majority of them concerned the English language, popularization of access to corpus tools and language resources, due to exponential progress of language technology software and the increasing availability of Internet connections, has enabled corpus compilations of several additional languages, many of them with no, or very little, (electronic) lexicographic tradition, e.g. Mirandese, Yoruba, and other Portuguese varieties (from Mozambique and Cape Verde).

Accordingly, irrespective of the theoretical paradigm from which corpus-based research is undertaken, it can be argued that corpus linguistics today owes its fundamental characteristic to Sinclair's ground-breaking demonstration of an unknown facet of the language: meaning derives from the relation that words maintain with other words.

3.1.2 Status of corpus linguistics

There has been considerable debate over the status of corpus linguistics, i.e. whether it is a theory, a methodology, an approach, a discipline or a branch of linguistics. A telling illustration of the lack of consensus are the answers provided by fourteen renowned corpus linguists associated with a wide spectrum of theoretical

[82] Here used as an umbrella term to refer to different phenomena ruled by the idiom principle.

frameworks, when asked to situate CL in the scientific or methodological panorama (Viana, Zyngier, & Barnbrook, 2011).

The vast majority of them, namely, Susan Conrad, Mark Davies, Stefan Th. Gries, Stig Johansson, Sara Laviosa, Geoffrey Sampson, Mike Scott, and John Swales, refer to CL with terms like 'methodology', 'approach', 'tools' or 'resources', in clear opposition to 'science'. For Guy Aston, Paul Baker, Ken Hyland, and Geoffrey Leech, however, CL is both science and methodology. A third position, held by Tony Berber Sardinha, states that the status of CL depends on the purpose of its use. Finally, Bill Louw affirms that CL is neither a science nor a methodology, but an instrument.

At this point, it should be mentioned that opinions diverge not only within the varied scope of linguistic perspectives interested in CL, as shown above, but also within one's own position over the course of time. One interesting example is Geoffrey Leech. While in 1992 Leech made sure to single CL out from, for example, sociolinguistics and psycholinguistics, as can be seen in this passage: "..is corpus linguistics really comparable with these other hyphenated branches of linguistics? No, because "corpus linguistics" refers not to a domain of study, but rather to a methodological basis for pursuing linguistic research" (Leech, 1992, p.105), twenty years later the researcher held a somewhat different point of view: "Corpus Linguistics is a methodologically-oriented branch of linguistics" (Viana et al., 2011, p. 158). This is a notable example of the depth and length of the debate.

Another view is that of Tognini-Bonelli (2001): CL as a *pre-application methodology*. For the author, "unlike other applications [of a methodology] that start by accepting certain facts as *given*, corpus linguistics is in a position to define its own sets of rules and pieces of knowledge *before* they are applied" (p. 1).[83]

Taken together, the diversity of opinions regarding the status of corpus linguistics indicates that CL is clearly a constantly developing object of interest.

[83] For further discussion on its status, see also, for instance, Sardinha (2000, pp. 355-357), Tognini-Bonelli (2001, pp. 1-2), and McEnery et al. (2006, pp. 7-8).

3.1.3 Applications of corpus linguistics

Nowadays, a variety of areas of linguistics utilizes corpora for their investigations.[84] Barlow (2011) has stated that corpus linguistics has contributed notably to the understanding of the nature of language in three spheres: provision of frequency information, quantification of the extent of variability in language, and emphasis on the importance of collocations.

These three scopes have been, to different extents, the foci of analysis not only of the corpus-oriented studies on academic language presented in Chapter 2, which include descriptions of language specialization (e.g., in terms of form, use, and function of lexical bundles) due to genre, area of knowledge, writing expertise status, language proficiency level, and register variation (see section 2.1), but also of other areas of language research that have used corpora, for instance, speech research, grammar, semantics, pragmatics and discourse analysis, sociolinguistics, stylistics, teaching of languages and linguistics, historical linguistics, dialectology and variation studies, psycholinguistics, cultural studies, and social psychology (see McEnery & Wilson, 1996, chapter 4, for a full account of these studies). Tognini-Bonelli (2001, p.1) also mentions gender studies and forensic linguistics, while Meyer (2002) adds translation and contrastive studies, natural language processing, and language acquisition.[85]

Particularly relevant for this present work is the notable contributions of the advent of corpora to the development of lexical studies, especially lexicography. Hence, section 3.2 below will present a brief overview of the history of the partnership between corpora and dictionary-making, while 3.3 will focus on corpora and corpus tools in the context of Portuguese and the use of corpora.

3.1.4 Corpus characteristics

Santos (1999) suggested the existence of two groups of people involved with corpus linguistics: the compilers, who were responsible for corpus design, review and annotation; and the users, those interested in retrievable information from the corpora.

[84] See the earlier mentioned book *Studies in Corpus Linguistics: Perspectives on Corpus Linguistics* (2011), edited by Viana, Zyngier and Barnbook for an elucidative overview of varied uptakes on crucial questions regarding corpus linguistics practice and theory.

[85] See chapter 1 of Meyer (2002) for a full account of corpus analysis and linguistic theory, and especially section 1.3 for corpus-based research in linguistics.

The author also observed the emergence of a third group, composed of corpus tool designers.

Almost 20 years later, one can see that, despite the continuing validity of the configuration of these three groups, at the same time boundaries have been increasingly blurred. On the one hand, easier access to sources of texts and development of user-friendly corpus compilation tools with no requirement of knowledge of programming (e.g., BootCat[86] for automatic crawling texts from the Web) have enabled people from the second group – corpus users – to create their own specially-tailored corpora. On the other, attempts have been made to facilitate the implementation of sophisticated features that further qualify the corpus by non-experts, for instance, GDEX configuration and Sketch Grammar devising in the Sketch Engine corpus tool.[87] Hence, a researcher can work on the three fronts, which makes corpus linguistics investigation increasingly more accessible.

3.1.4.1 Corpus design

Corpus linguistics methodology (Biber, Conrad, & Reppen, 1998; McEnery et al., 2006; Sinclair, 2003a, 2003b, 2005) as explained in section 3.1 above indicates that these are the factors to be considered when building a corpus: corpus size; balance and representativeness; data capture; corpus mark-up; corpus annotation; and character encoding (McEnery et al. 2006, pp. 71-76).

Definition of the corpus size may follow two perspectives. One option is to determine the total number of tokens required for the corpus to be used for purpose **X.** So, for instance, a decision was made that corpora of Portuguese varieties should each total at least 30 million tokens (Almeida, Ferreira, Correia, & Oliveira, 2013) in the project that created the *Vocabulário Comum da Língua Portuguesa* ('Common Orthographic Vocabulary of the Portuguese Language') (VOC) (see Chapter 1).

[86] http://bootcat.dipintra.it/

[87] Although handling these functions requires familiarity with the corpus tool and dedication to learn specific skills, it is not necessary to be a computational linguistic or a language engineer. I am an example of such a researcher. The Sketch Engine corpus tool will be presented in Chapter 5, Corpus creation in Chapter 6, Sketch Grammar for academic Portuguese in Chapter 7, and GDEX configurations in Chapter 8.

Another method consists of selecting everything that is available on the stretch of language of interest, with a final total size corresponding to what was actually available. This was the approach taken in this thesis, as will be shown in Chapter 6.

Corpus balance refers to the distribution of texts according to the kind of texts comprising the corpus. In a multi-genre corpus, for instance, it might be decided that distribution of texts should be balanced among them. Another possibility, in corpora covering various language varieties, is to determine an equal number of token per language variety subcorpora. This is the case of the corpus built for this thesis (see Chapter 6).

Representativeness concerns selecting texts samples that represent the stretch of language under study. Taking the BNC corpus[88] as an example, it aimed at representing contemporary English language as used in Britain. Thus, samples of texts were collected from academic journals, newspapers, magazines, literature books, brochures, among others, to represent written English. For spoken English, many hours of interviews, phone conversations, lectures, etc. were transcribed and included in the corpus. Although the BNC is not balanced (e.g., only 10% of which are oral texts), it has certainly tried to be representative.

Following this first stage, and to prepare for the next one (corpus annotation), the corpus needs to be automatically and/or manually cleaned up. The extent of elements to be cleaned, once again, depends on the purpose of the corpus being designed. For lexicographical purposes such as dictionary-making, it is common to keep only textual information, so figures, table, charts are usually removed.

It is advisable to work on character encoding before the clean up since any problematic transformation can be spotted and, if pervasive, encoding conversion can be reviewed. Nowadays, UTF-8 encoding has been widely used in a series of corpus tools.

The next phase is corpus mark-up and entails adding marks to texts in order to contribute to advanced corpus analysis. Mostly, metadata are included in texts through headers. Totality and the kind of information contained in headers depend on how many external data were recorded when collecting texts. For instance, in the specially built

[88] http://www.natcorp.ox.ac.uk/]

corpus for this thesis, texts were marked with area of knowledge, year of publication, ISSN number, to name but a few.

Finally, corpus annotation refers to adding labels to every word in the corpus in order to allow several ways of corpus analysis. This phase consists of a number of processes of identification and transformation; nowadays, many annotation tools accumulate all these roles into one.

A tokeniser goes through all strings of characters and separates them in tokens, where one token corresponds to any string of characters between two spaces. This means that tokens comprise words and punctuation.

Another tool – the lemmatiser – analyses the tokens, identifies their form and turns them into citation forms of words when necessary. For instance, the token *estudos* ('studies', noun) is in the plural form; the lemmatiser transforms it in *estudo*. The same applies to inflected verbs and adjectives, the former being converted into the infinitive form, while the latter is presented in the masculine and singular forms.

POS-taggers are tools that identify tokens and assign part-of-speech tags to them. Attention should be drawn to the fact that these tags not only inform the word class, e.g., noun, verb, adverb, etc., but also show the kind of inflection of that particular word. Tags take the form of codes and, to interpret them, lexicographers must consult the tagset of the tagger used for corpus annotation.

Alternatively, assignment of tags according to syntactic relations are also used; the tool employed is a parser. A major advantage of this kind of annotation is that performance is very high, meaning that tags are correctly annotated with POS and also syntactic function in the vast majority of cases. The downside is that parsers are only available for a few languages. Moreover, it has been shown that commercial tools perform much better than free tools, which is another shortcoming as paid parsers can be very expensive in the context of academic projects.

3.1.4.2 Corpus analysis: corpus-based X corpus-driven approach

According to corpus linguistics literature, there are two approaches for corpus analysis: *corpus-based* research and *corpus-driven* research. Their principles sometimes overlap, and sometimes are diametrically opposed, depending on the author's point of view (see Biber et al., 1998 for a perspective on corpus-based research; Tognini-

Bonelli, 2001, for an account of corpus-driven approach; Hanks, 2012b and McEnery et al., 2006 for a discussion on the differences between them). In fact, McEnery et al. argued that "the distinction between corpus-based vs. corpus-driven approaches is overstated" (2006, p. 8).

In the tradition that opposes these two methods, a *corpus-based* approach "seeks to support preconceived theories with judiciously selected examples from a corpus" (Hanks, 2012b, p. 417). Here Hanks criticised linguists who develop language theories without analysing actual language use, turning to corpora only in search for examples that can support their theories.

This perspective can be extended to lexicographic projects in which corpora are used exclusively as sources for examples. In other words, corpus-based dictionaries that follow this approach have not engaged in analysis of language behaviour in the corpus.

The corpus-driven approach, on the other hand, describes the lexicogrammatical characteristics of a lexical unit based solely on corpus data. As originally conceptualised (e.g., Tognini-Bonelli's), this approach opposes the use of annotation in corpora. In this view, the starting point of analysis should be the concordance lines that were produced to a certain word form. Thorough manual evaluation of all lines then indicates word classes of the keyword, and sorting lines according to neighbouring words leads to sense differentiation, identification of collocations and syntactic and semantic patterns. The argument foregrounding this perspective is that corpus annotation "taints" the corpus by pre-establishing elements of the analysis.

Although the justification is sensible, the fact is that lexicographic work nowadays deals with gigantic corpora, so it is virtually impossible for lexicographers to go through every concordance line and manually analyse a word's behaviour.

Despite this clear-cut distinction between the two types of corpus analysis approach, it is widely known that, nowadays, both terms are used interchangeably (at least in the context of dictionary writing) to designate performance of language description informed by corpus data.

3.2 Corpora and lexicography: the electronic revolution

The advent of electronic corpora has opened up unparalleled opportunities to lexical analysis, leading to an unprecedented revolution in lexicography (cf. Hanks, 2012a, 2012b; McEnery et al. 2006; Rundell, 1998). If in the past corpora were composed of handwritten citation slips (e.g., for the production of the famous Oxford English Dictionary) and "most dictionaries…were based on a mixture of citations, introspection, and what other dictionaries said" (Moon, 2007, p. 165), now electronic corpora have enabled major transformations in the process of dictionary-making. Some of the new possibilities are (Biber et al., 1998, p. 1; Hanks, 2009; McEnery et al., 2006, pp. 80-81):

- presentation of authentic, typical examples of the usage of a lexical item;
- recording of frequency information;
- information on phraseology;
- provision of frequent collocations;
- more sensible grouping of words which are polysemous and homographs;
- more accurate dictionary entries;
- information on contexts of use;
- account of syntactic patterns (grammar);
- informed decisions on what to leave out (coverage);
- description of (conversational) pragmatics;

Taken together, these tasks are part and parcel of the process of dictionary-making. Rundell and Kilgarriff (2011) pointed out a series of actions required for creating a dictionary from scratch (see Table 3.1):

Table 3.1 Inventory of the main actions for dictionary creation (Source: Rundell and Kilgarrif, 2011, p. 261)

- corpus creation - headword list development - analysis of the corpus: - to discover word senses and other lexical units (fixed phrases, phrasal verbs, compounds, etc.) - to identify the salient features of each of these lexical units 1. their syntactic behaviour 2. the collocations they participate in 3. their colligational preferences 4. any preferences they have for particular text-types or domains - providing definitions (or translations) at relevant points - exemplifying relevant features with material gleaned from the corpus - editing compiled text in order to control quality and ensure consistent adherence to agreed style policies.

At this point, it should be noted that this exceptional trend – corpora use in lexicography – has gained momentum due to tremendous improvements of Natural Language Processing (henceforth NLP) technologies, also known as computational linguistics, language engineering, human language technologies (HLT) (Kilgarriff & Rundell, 2002, p. 812) . This is a key issue since, as Rundell (2009) affirmed, what seems to have changed in modern lexicography is the balance between computer use and the lexicographer's routine job, with the former increasingly taking on the role of humans. Given this, the author moves on to present two aspects to be taken into consideration for reflection about future developments in the field:

- technologies that have enabled us to do the same things we did before, but more efficiently and more systematically.
- "game-changing" developments that have expanded the scope of what dictionaries can do and (in some respects) changed our view of what dictionaries are for (Rundell, 2009, p. 9).

Drawing on these two pointers, I will here provide a succinct birds-eye-view of the progress of corpora use in lexicography[89], focusing on the growing transfer of lexicographers' tasks to the computer.

[89] Two preliminary remarks should be made here concerning some limitations of the present overview. Firstly, most of the discussion presented in this section is in reference to the Anglophone tradition. This is due to long-standing English-oriented development of new technologies for language research and the subsequent application of the outcomes of this research to other languages. Thus, this review will

3.2.1 A revolutionary dictionary

Rundell and Kilgarriff (2011, p. 259) pointed out 1981 as the "Year Zero" of modern lexicography, with the beginning of the COBUILD project, which set out to produce the Collins COBUILD English Language Dictionary (Sinclair & Hanks, 1987) for advanced learners of English.[90] The role of the corpus in this dictionary was overriding: "For the first time, a large-scale description of English was created from scratch to reflect actual usage as illustrated in (what was then) a large and varied corpus of texts" (Rundell and Kilgarriff, 2011, p. 229). Hanks (2009, p. 216) informed that lexicographers initially worked on a corpus of 7.3 million words – which was seven times bigger than any other previous corpus, thus impressive for the time – to move on to an enlarged 18 million-word corpus[91] as the project headed to its end, and could find uses and regularities of the English language never seen before.

Jeremy Clear, who was the Senior Computer Officer in the COBUILD project, has given a detailed account of all the technological innovations to lexicography addressed by making the Collins COBUILD dictionary (Clear, 1987, pp.41-61). He highlighted the fact that behind these extraordinary discoveries was a set of project-tailored new computational programs.

Clear starts by stating that KWIC concordances, word lists builders and frequency counters, all already existing programs for text processing, had to go through considerable adaptations in order to attend to the demands of the COBUILD project, thus resulting in the creation of new resources for corpus processing. He goes on to explain the special building of a unique lexical database to record data related to each headword, including definition, inflected form, syntax, semantic field, style, synonyms, collocates, and examples – the latter being copied word for word from the text in the

probably either refer to the reality of other languages as well, or anticipate new trends. Nevertheless, it should be stressed that corpora in lexicography of the Portuguese language is reviewed separately in the next section. Secondly, it has been shown that the vast majority of the innovations in lexicography have originated from projects creating monolingual learners' dictionaries of English (MLD) (see Rundell, 1998). This is not surprising, given the huge profit that the MLDs business generated. Constant integration of innovations were efficient marketing strategies for publishers to find (or keep) their position in the fiercely competitive market of the time. However, monolingual, bilingual, and specialized dictionaries in a variety of languages have also benefitted from the advent of corpora. For a further account on this particular aspect of the corpus revolution in lexicography, see Hanks, 2009, 2012b).

[90] See Sinclair (1987) for a detailed account of the COBUILD project.

[91] These corpora were versions of the Birmingham Collection of English Text (BCET), which later became the Bank of English (Hanks, 2012b, p. 412).

corpus (Clear, 1987, p. 42). The author also informs that application software was created to handle the data in the database. One of them, *update dic*, enabled editing of entries displayed on the computer screen through the use of cursor-control keys. Another allowed automatic cross-reference checking and thus was deemed particularly valuable due to the elimination of some common inconsistencies usually found in dictionaries that did not use this program. Finally, Clear points out the creation of a program for Dictionary Extraction and Editing (pp. 59-61) employed in the final stage of the project, i.e. dictionary publication.

Here an absolutely innovative feature integrated into the above-mentioned program should be highlighted that automatically converted originally manually-written definitions into the pioneering COBUILD-style.[92] The computer would provide a definition and the lexicographer only had to verify the entry already in the database and edit it only when necessary. The program "used the syntax information associated with each sense category to generate an appropriate phraseology for the formulaic beginning of each definition" (Clear, 1987, p. 59). It is particularly interesting that, despite some wrongly-adapted definitions, the overall conclusion was that such automation had significantly facilitated lexicographers' work (Clear, 1987, p. 59). The principle underlying this innovation is at the heart of some present dictionary-making projects (DOPU being one of them), as it will be shown below.

As can be seen, the use of electronic corpora and computational technology in the COBUILD project has dramatically changed the role of lexicographers. Firstly, they were no longer using intuition for entry writing (i.e. definition, description of syntagmatic patterns, collocations, examples, synonyms, antonyms), but basing this key lexicographical task on the information obtained from authentic evidence of language use in the corpus. Secondly, special corpus analysis tools facilitated the lexicographer's work tremendously, i.e. concordancing and sorting allowed for precise headword meaning apprehension, identification of collocations, and description of patterns; and word frequency counts permitted a clear overview of words' typicality, also resulting in real use-based information for labels. However, as Clear reported, "there was no

[92] Also known as "full-sentence definition", it consists of explaining the meaning of a word by employing it in the explanatory definition. For instance, one sense of the verb 'listen', in the COBUILD dictionary, is "If you *listen* to someone who is talking or to a sound that you can hear, you give your attention to them or to it" (Hanks, 1987, p. 122). See Chapter 11 for an account of definition strategies.

possibility of extracting the linguistic information required for the dictionary from the corpus automatically" leading to still manual production of traditional lexical analysis, namely, "definition, grammatical information, semantic field labels, synonyms, etc.".

3.2.2 Larger corpora, more resources, better dictionaries

It can be said that such a successful corpus-driven approach to making the Collins-COBUILD dictionary set the basis for corpora design and analysis, language description, and also dictionary-making, radically revolutionising lexicography.

In this vein, other dictionaries of English for learners followed suit, taking advantage of ever growing corpora and rapid advancement of corpus-oriented technologies. For instance, Longman Dictionaries (Summers, 1993) specially compiled the 28 million-token Longman/Lancaster English Language Corpus for creating the Longman Dictionary of Contemporary English (LDOCE). This corpus was incorporated into the 100-million-word British National Corpus (BNC), which in turn was employed for the creation of the third edition of LDOCE (1995) and the Oxford Advanced Learner's Dictionary of Current English (OALD) (6^{th} edition, Wehmeier, 2000) (cf. Hanks, 2012b). As a notably large corpus for its time with a carefully-designed compilation process (cf. Atkins, Clear, & Ostler, 1991), the BNC marked another revolution: now corpora sizes were heading up over 100 million tokens.

With regards to technological progress, many innovations in the 1980s and 1990s have substantially contributed to enhancing efficiency and systematization of lexicographic working routines. Rundell and Kilgarriff (2011, pp. 259-260) highlighted the fundamental importance of the upsurge of corpus tools like tokenisers, lemmatisers and POS-taggers in enabling more sophisticated language analysis, while the creation of dictionary writing systems introduced consistency in entry styles as lexicographers were presented with a set of predefined options for certain data fields, eliminating errors usually made by humans.

Whereas it is true that the great volume of data resulting from the use of bigger corpora have enabled much richer language description - thus more accurate dictionary features - it cannot be forgotten that tools for corpus analysis had not advanced at the same pace, so access to so much data started to pose some serious challenges to

lexicographers. For instance, it has been shown that the number of concordance lines had grown to unmanageable lengths for humans (Kilgarriff, 2003).

An alternative to tackling this problem was then put forward: the provision of statistical summaries. Church and Hanks (1990) proposed the Mutual Information (MI) association measure, which enabled identification of collocations and word meaning analysis. Since then, many other statistics have been created (e.g., MI3, log-likelihood, logDice), and it can be said that adoption of statistical measures for words co-occurrences analysis has become a routine procedure in dictionary-making projects (Kilgarriff, 2003).

Kilgarriff (2003, s.p.) has claimed that, even though statistical summaries have contributed greatly to organising the presentation of information, thus facilitating lexicographers' job and enhancing language analysis, "they are not used as widely as they might be". The author went on to point out three weak aspects of (at that time)[93] current use of statistical summaries from a lexicographic standpoint: the statistics used were not ideal: too much noise, made up of uninteresting words found in the neighbourhood of the node; and the fact that neighbourhood refers to span window searches for collocates to the left and to the right of the node, irrespective of the grammatical function they reflect. This is a key point as it demonstrates that, although much improvement has been achieved, "first-generation summaries mix everything together, so we have to sift through objects, modifiers, pronouns, proper names, adverbs and everything else" (Kilgarriff, 2003, s.p.).

3.2.3 Automation begins

For Kilgarriff (2003), one possible solution to this problem was new software that automatically identified collocates in very large corpora and organised them according to the grammatical relations maintained with the headword, yielding lexical profiles for each searched word. A program called word sketch (Kilgarriff & Tugwell, 2001), which had been originally created to provide input for a system handling Word Sense Disambiguation (see Kilgarriff & Tugwell, 2001), provided these pre-sorted word summaries and seemed a very useful alternative.

[93] In fact, many statistical measures created at that time are still being widely used in corpus-based studies, including in lexicography.

As Rundell and Kilgarriff (2002) reported, given the potential benefits that the use of a summary of a word's behaviour could confer to dictionary-making, both in terms of dictionary-completeness enhancement and cost reduction, lexicographers working on the creation of the *Macmillan English Dictionary for Advanced Learners* (MEDAL) used word sketches for over 8,000 words. Kilgarriff and Rundell pointed out that, indeed, lexicographers' work was highly facilitated by having a summary of a word's collocational behaviour at hand to complement the usual process of concordance line analysis, besides contributing to the difficult task of sense differentiation.

However, as "Word Sketches came to be the lexicographer's preferred starting point for analysing a given word" (Rundell & Kilgarrriff, 2002, p. 817), that is, "they [lexicographers] used the word sketch as the first and main view of the corpus data, with KWIC concordances only being used where there was some issue needing further investigation" (Kilgarriff, 2003, s.p.), the authors soon realised that the word sketches had played an even more important role: they led to significant changes in corpus analysis methodology. In other words, the initial stage of corpus analysis was automatically performed by the computer, which then provided the lexicographer with pre-digested information on a series of aspects (as mentioned earlier).

With the benefit of hindsight, it can be argued that this was a "game-changing" development in lexicography. As Rundell and Kilgarriff pointed out, while until 1997 (year of the initial planning for the earlier mentioned MEDAL) the computer's main role was to facilitate lexicographers' work, from that moment on, "some of the key lexicographic tasks [were] beginning to be transferred, to a significant degree, from humans to machines" (Rundell and Kilgarriff, 2011, p.257).

As computers took on the job of identifying the grammatical relations in which a word participates, separating them to later present the results in organised boxes with lists of collocates ordered by frequency and salience, it is not surprising that this method has proven to reduce dramatically the lexicographers' work, hence increasing efficiency and cutting costs, as well as systematizing the process of language description, making it more reliable (Rundell and Kilgarriff, 2011, p. 257).

It is unquestionable that these innovations reflect the overwhelming effects of technology, and more specifically, the Internet, on various fields of language research,

which are required to rapidly adapt or else run the risk of obsolescence. As expected, lexicography has inevitably encountered this situation, and important steps have been taken in order to attend to this new demand and keep up with modern trends.

3.2.4 e-Lexicography

Reflecting the dramatic impact of the technological revolution in lexicography, a subfield has arisen that is referred to as electronic lexicography, or e-lexicography for short. Sylviane Granger explains that the use of the term *electronic lexicography* in the seminal book with the same name that she co-edited with Magali Paquot (Granger & Paquot, 2012) concerns:

> (…) the design, use and application of electronic dictionaries (EDs), which are in turn defined as primarily human-oriented collections of structured electronic data that give information about the form, meaning and use of words in one or more languages and are stored in a range of devices (PC, internet, mobile devices) (Granger, 2012, p. 2).

According to Granger, while some current lexicographic projects only make superficial use of the radical possibilities offered by electronic media, for instance, by simply providing e-versions of printed dictionaries (which seems to be the case of many dictionaries of Portuguese; see next section), electronic lexicography goes far beyond a mere change of format to incorporate a series of cutting-edge innovations. Table 3.2 summarises her view, including some opportunities and challenges considered to accompany these changes:

Table 3.2 Changes in lexicography due to computational technology. (Source: Granger, 2012, pp. 3-5)

Innovation	Opportunities	Challenges
Corpus integration	Production of rich lexical entries; Part of the dictionary for users to access	No longer raw data as primary resource, but rather pre-analysed and pre-sorted data
More and better data	Free space allows for presentation of richer collocational coverage, more examples, multimedia content and extended notes	Users "swamped" with data; Space restrictions in small-screen devices
Efficiency of access	Wide range of search options; Navigations within and beyond the dictionary due to hyperlinks	Access still not optimal
Customization	Adaptable dictionaries: manual customization by the user; Adaptive dictionaries: automatic adaptation to users' needs	Research on users' needs concerning electronic dictionaries
Hybridization	Combination of one or more types of reference work in a single product	Dictionary as an integrated tool
User input	Increased number of entry writers; Up-to-date language change and lexical innovation	Accuracy is questionable

As can be seen, while innovations are undoubtedly advantageous to lexicographers and users, they also pose some completely new difficulties, which were unheard of in traditional lexicography, thus requiring appropriate handling.

Some measures that have been taken to foster productive dialogue, knowledge exchange and the building of collaborative multidisciplinary expertise include, for instance, the establishment of a new biannual conference specifically focused on e-lexicography. The *electronic lexicography in the 21st century (e-Lex)* conferences[94] were inaugurated in 2009, in Louvain (Belgium), and are now heading to their fifth edition in September of this year, in Leiden, the Netherlands. The cross-disciplinary aspect of the conference is highlighted in its presentation on its website:

> Electronic lexicography in the 21st century (eLex) conferences aim to explore innovative developments in the field of lexicography. We are in an interdisciplinary field, and eLex brings together specialists in dictionary publishing, corpus lexicography, software development, language technology, language learning and teaching, translations studies and theoretical and applied linguistics.

[94] https://elex.link/

Another notable initiative is the *European Network of e-Lexicography (ENeL)*, which is an action under the European Cooperation in Science and Technology (COST) framework that aims 'to increase, co-ordinate and harmonise European research in the field of e-lexicography and to make authoritative information on the languages of Europe easily accessible'.[95] It is composed of four working groups with different objectives and interests, which reflect the wide-reaching scope of the action: WG1: integrated interface to European dictionary content; WG2: retro-digitized dictionaries; WG3: innovative e-dictionaries; and WG4: lexicography and lexicology from a pan-European perspective. The action promotes regular meetings on specific predefined topics, training schools, workshops and short-term scientific missions.[96] It started in October 2013 and will finish in the same month in 2017.

From the perspective of the two criteria adopted for reviewing the relationship between corpora and lexicography as exposed at the outset of this section, it is unquestionable that we are experiencing a new, profoundly impactful, game-changing development brought forth by the advances in e-lexicography as a new field. Without a doubt, definitions of what dictionaries are and what they are for no longer fit traditional categories and taxonomies.

3.2.5 A new era

E-lexicography as a well-established field is, at the same time, the result and the promoter of the production of new tools and resources focusing on enhancements in corpus data creation and interrogation, as well as alternative methods that streamline lexicographic work.

One example of a growing practice in e-lexicography is the automatic acquisition of knowledge for lexicographical projects, as shown in the survey conducted by Tiberius, Heylen and Krek (2015) with the 129 members of Working Group 3 (Innovative e-Dictionaries) of the European Network of e-Lexicography (ENeL).

[95] http://www.elexicography.eu/

[96] As a member of working group 3, I had the unparalleled opportunity to participate in many of these activities, including meetings, workshops, a training school on tools and methods for creating innovative e-dictionaries (Ljubljana, 2016) and a Short-term Scientific Mission with Dr. Iztok Kosem at the University of Ljubljana (Slovenia). I am deeply grateful to Dr. Robert Lew, who informed me about the action.

Respondents were asked to indicate, in the context of their projects, if, and if so, which lexicographic tasks were automated. As the graphic (Figure 3-1) below shows, the top two types of automatically acquired knowledge are lemma list and frequency information, followed closely by examples.

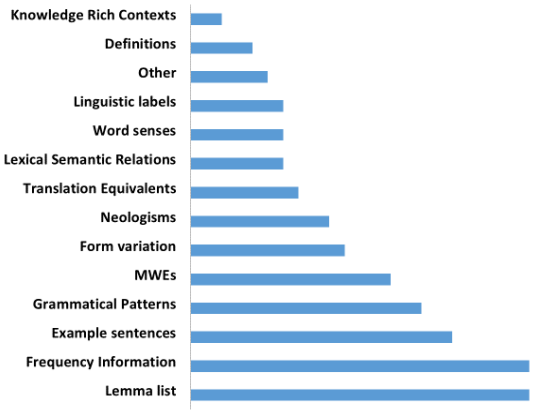


Figure 3-1 Automated lexicographic tasks in projects across Europe (Tiberius, Heylen, & Krek, 2015)

Of 50 valid answers, 36 indicated use of some form of automatic acquisition of knowledge. These answers referred to lexicographical projects in the following countries: Basque Country, Belgium, Bulgaria, Czech Republic, Denmark, Estonia, France, Germany, Greece, Hungary, Italy, Netherlands, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden and Switzerland.

Notwithstanding the non-exhaustive nature of this survey, these results suggest that the focus of initially English-centred technological advances has widened to reach many other languages.

Another important characteristic of present-day e-lexicography is the possibility of making use of huge corpora, which are reaching sizes as large as over 19 billion words (e.g. TenTen[97] corpus family at the Sketch Engine) due to highly advanced tools for crawling Internet pages. As mentioned earlier (see section 3.1.3), the interest in using the web as either a corpus itself or as a source for corpus compilation, together with the necessary development of technology to achieve these goals, is not new,

[97] "TenTen is a new generation of Web corpora. These corpora are created by Web crawling and processed with our latest boilerplate cleaning and de-duplication tools. The "TenTen-corpus" designates the target sizes of the corpora which is 10'10 (10 billion) words" (text informed on the Sketch Engine website). For more information, see Jakubíček, Kilgarriff, Kovář, Rychlý and Suchomel (2013).

though, having begun over a decade ago.[98] What is especially notable about the current status of research in this area is that it is very sophisticated, resulting in more qualified (much less junk due to boilerplate removal, no text duplicates), very large corpora, which, accordingly, become a trustworthy source for dictionary creation.

It has been argued that the greatest advantage of using these gigantic web corpora for dictionary-making is that "for the computer, the bigger the corpus, the better the analyses: the more data there is, the better the prospects for finding all salient patterns and for distinguishing signal from noise" (Rundell and Kilgarriff, 2011, p. 269). Nevertheless, this is not a consensual position; opponents claim that fine-grained descriptions of language use that take into consideration variables like text genre, time, size, place of publication, language variety, besides other information of a more sociolinguistic level (like author's age, gender, formal education, etc.), cannot be done based on corpora with very limited mark-up (it is usually possible to narrow down corpus search only to aspects related to the Internet, like domain (.edu, .com, etc) or URL address).[99] In any event, as has been shown earlier (see section 3.1.2), the purpose for using corpora determines which size is best suited for undertaking a specific language search, and thus, ultimately, the user should make the decision.

Nevertheless, it has been claimed that very large corpora are only beneficial to language studies if all the richness of information that they contain can be properly made visible to researchers. Hence, some data pre-treatment is crucial, especially as manual inspection is virtually impossible. That means that analysis optimization of these mega-sized corpora requires the most advanced corpus query tools, and demands even more automation of lexicographic work.

One of the proposals for further automating lexicographic work has been put forward by a team of researchers in Slovenia, who developed "the semi-automated approach" for dictionary-making (see Gantar et al. 2016; Kosem et al. 2013; Kosem et

[98] See Biewer, Nesselhauf, and Hundt (2007) for a full account of the initial discussions of this area of interest. In addition, consult the website of *the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus* (ACL SIGWAC), which was founded in 2005.

[99] The PtTenTen Web 2011 Palavras is a web-crawled corpus of Portuguese which presents information on language variety, too. This information was provided due to a request from the Oxford University Press, which needed to cover both standard varieties of Portuguese in order to create a bilingual dictionary of Portuguese – English, called the Oxford Dictionary of Portuguese (Frankenberg-Garcia & Newstad (Eds), 2015).

al. 2014; Logar & Kosem 2013), drawing on Rundell and Kilgarriff's (2011, p. 278) original idea that envisaged:

> …a change from the current situation, where the corpus software (some version of the word sketches) presents data to the lexicographer in …intelligently pre-digested form, to a new paradigm where the software selects what it believes to be relevant data and actually populates the appropriate fields in the dictionary database.

According to the authors, time-consuming activities of manual selection and copying of information from the corpus tool to dictionary writing system are substituted for validation of the extracted data directly in a structured database.

In this vein, the above mentioned semi-automated approach implements Rundell and Kilgarriff's revolutionary view by employing specially-tailored, high-end technology for automatic extraction of data from corpus and import into dictionary writing systems.

As described in Gantar et al. (2016), word sketches are in the heart of this radically innovative method consisting of automatically extraction from the corpus, for each keyword, of information on grammatical relations, collocates, collocates frequency and salience score, and a certain number of examples per collocate and import, via API script, to the dictionary writing system, where then the lexicographers' job is narrowed down to validation, analysis and edition of the entry.

A key consequence of the adoption of this approach is that the provision of pre-digested information leads to considerable streamline of the lexicographic process. Furthermore, and more importantly, the transfer of lexicographic tasks from the lexicographer to the computer has proven to not taint the quality of final entry's, as information is not lost. This conclusion has been reached from the undertaking of a carefully structured experiment in which manual approach was compared to semi-automatic approach. The results indicated that the method not only effectively works, but is also applicable to other languages.

Thus, given the consideration of the semi-automated approach as the state-of-the-art method currently available for dictionary-making, it was my decision to adopt it for creating the design of DOPU. The approach will be further touched upon in Chapter

4, while the entire part II of this thesis is devoted to the presentation of the development of the requirements for implementation of this method for DOPU creation.

3.3 Corpora, NLP tools and Lexicography of the Portuguese language

Without a doubt, the level of success of corpus-driven dictionary-making projects relies greatly on the kind and quality of electronic resources employed. It is thus not surprising that very good dictionaries of English have been made since the ground-breaking corpus revolution in lexicography, which was only possible due to the accompanying progress of computational linguistics (see section 3.2 above).

Electronic corpora and computational tools have been utilised in research on Portuguese since the beginnings of the digital revolution in language investigations (see below). It should be noted, however, that it was not until the 1990s that Brazil and Portugal witnessed a significant growth in the development of computer-based data and resources for language treatment.

Considered the researcher responsible for introducing lexical-statistical analysis of Portuguese to the Brazilian linguistics community as early as the 1960s, Professor Maria Tereza Camargo Biderman was one of the pioneers in the compilation of electronic corpora and use of computational tools in Brazil - at a time when the term corpus linguistics was not even being used yet.

According to Biderman (1978, pp. 65-67), a review of some of the inaugural studies undertaking a language-technology-based analysis of Portuguese pointed out investigations of both Brazilian and European Portuguese language varieties in the United States, France and Brazil. The author also observed three main lines of computer-based research regarding Portuguese at that time: for literary and/or stylistic purposes; for linguistics purposes; and for computational purposes (Biderman, 1978, p. 64). These pioneering works are summarised in Table 3.3.

Table 3.3 Pioneering work on computer-based lexical statistics. (Source: Biderman, 1978, pp. 64-67)

Place	Study	Author	Purpose
Centre of Lexicology and Stylistics of the Spanish and Portuguese languages[100] at the University of Toulouse (France). Coordination: prof. Jean Roche	Vocabulary and Lexical Analysis of *Auto da Compadecida*[101] – PhD thesis (published in Brazil in 1969).	Jacques Emorine	Literary/stylistics.
University of São Paulo (Brazil)	Computational Analysis of Fernando Pessoa (An Essay of Lexical Statistics) [102]. PhD thesis. 1969.	Maria Tereza Camargo Biderman	Literary/stylistics.
University of Georgetown (USA)	Syntax of Portuguese. PhD thesis.	Cléa Rameh	Linguistics
Stanford University (USA)	Dictionary of Frequencies of written European Portuguese. PhD thesis. 1972.	John Duncan	Frequency/dictionary. Linguistics.
***Naval Academy of Anápolis* (Brazil)**	Concordances (KWIC) of each word extracted from a 400,000 word corpus of spoken Brazilian Portuguese. Use of IBM Magnetic Seletric Typewriter MT/ST (optical reader).	Prof. J. Hutchins and collaborators	Spoken Portuguese/ concordance. Linguistics.
University of São Paulo (Brazil)	The Gender Category[103]. Professorship thesis. 1974. Lexical analysis of romance language dictionaries of frequency.	Maria Tereza Camargo Biderman	Frequency/dictionary. Linguistics.
Division of Electronic Engineering at ITA – Institute of Aeronautics Technology. (São José dos Campos, Brazil)	Determination of entropy in the Portuguese language	O.A. Nawa (1965), A. Schoenacker Filho & Paulo de Tarso Ribeiro (1966), M.S. Berman (1967) and P.A. Abreu & R. Nascimento Melo (1968)	Computational.
Division of Electronic Engineering at ITA- Institute of Aeronautics Technology. (São José dos Campos, Brazil)	Viability of Literary Studies via Computer. Compilation of a corpus comprising novels written by Brazilian authors, from different literary styles and times. 1968.	Four engineers.	Computational.

Although not mentioned in Biderman's brief review, to this list of pioneering computational corpus-based works should be added the previously presented (section 2.2 above) *500 hundred Brazilian Portuguese Word List* created by John R. Kelly in 1970, which was extracted from a 128,000-word electronic corpus especially compiled for this purpose.

100 Centre de Lexicologie et Stylistique des Langues Espagnoles et Portugaise.

101 Lexique et Analyse Lexicale de L'*Auto da Compadecida*.

102 *Análise Computacional de Fernando Pessoa* (*Ensaio de Estatística Léxica*).

103 *A Categoria do Gênero*.

Biderman was also a pioneer in corpus-based dictionary-making in Brazil. As early as 1984, she published a seminal paper[104] on how to compile a dictionary of general Portuguese. There, the author defended that a large multi-textual-genre corpus should be compiled for this purpose, suggesting corpus frequency as the criterion for headword selection. She called attention to the effects on nomenclature according to the decisions taken for tackling the issues of single versus complex lexemes and homonym versus polysemy. She was also fully aware of the use of controlled vocabulary for definition writing, praising the innovative work of the Longman Dictionary of Contemporary English (1978). Her innovative corpus-based *Dicionário de Português Contemporâneo* (1992), aimed at elementary school students, was the first lexicographic-sound dictionary published in Brazil (Krieger, 2015).

In Portugal, the progress of language technology was also playing a revolutionary role in the investigation of the Portuguese language. In the 1970s, the adoption of a corpus-driven approach using computational tools for the compilation of a corpus of spoken European Portuguese granted much more sophistication and rigorous analysis for the creation of the *Português Fundamental* lemma list[105] in comparison to the inaugural project, developed in France between 1951 and 1953, when the corpus for the *Français Fondamental 1er degré* [106] was manually collected and analysed (Rivenc, 1996, p.23).

It should be mentioned that this vocabulary was the basis for Professor Mário Vilela's *Dicionário do Português Básico* (1990), which was targeted at learners of Portuguese, both speakers of PAL and Portuguese as a mother tongue. It was considered a revolutionary dictionary due to the employment of a non-traditional microstructure, one that privileged the presentation of word senses in terms of their uses, i.e. with a

[104] *O dicionário padrão de língua* (Biderman, 1984b).

[105] It should be mentioned that Biderman's work on the Fundamental Vocabulary of Brazilian Portuguese (see, for instance, Biderman, 1996a) was inspired by this project.

[106] According to Rivenc (1996), the purpose of this innovative project was to create pedagogical materials, namely, vocabulary and grammar, for teaching languages (French, Portuguese, among others) to foreign students, based on real language use evidence. The Portuguese branch of the project is notable due to the use of computers for corpus compilation and analysis. The results have been published in Nascimento, Marques and Cruz (1987), and Nascimento, Rivenc and Cruz (1987). For a detailed and commented account of the project, together with an attempt to propose a fundamental vocabulary of Brazilian Portuguese, based on the original European version, see Biderman 1996a, 1996b.

More information is available at http://www.clul.ulisboa.pt/en/11-resources/278-spoken-corpus-portugues-fundamental-pf.

focus on collocational and colligational patterns grounded in the valency grammar approach. Notwithstanding its lexicographic-rich structure, the *Dicionário de Português Básico* was not particularly user-friendly, requiring advanced dictionary use skills.

A steep increase in the development of natural language processing methods and tools marked the Portuguese language research scenario beginning in the 1980s, with considerable intensification in the 1990s, most notably in Portugal due to significant incentives from the European Union. Aimed at facilitating communication in a multilingual Europe while fostering technological means for the preservation of local linguistic identities, the European Community funded a series of projects and programmes that levelled Portuguese with other European languages.

In 1987, Portugal embarked on the ambitious 10-year project on automatic translation called EUROTRA (which had begun in 1982 in other countries), hosted by the *Instituto de Linguística Teórica e Computacional* (ILTEC)[107] ('Institute of Theoretical and Computational Linguistics'), which had been created specially to undertake this project (Branco et al., 2012, pp. 65-66). Mateus and Branco (1995) commented on the benefits of the participation of Portugal in European-wide initiatives to the development of computational linguistics for Portuguese:

> [A]dditional projects have been launched aimed at the construction of electronic grammars and dictionaries, as well as interfaces in natural language; the creation of terminological databases and electronic corpora, spell checkers and morphological analysers; and also, the creation of resources for educational technology. As can be seen, Portuguese is at the same level as other major European languages.[108] (Mateus & Branco, 1995, p. 8) (my translation).

Regarding these projects, Correia (1994) gave an instructional account of the state of computational lexicography in Europe in the first half of the 1990s, with a special focus on digital lexical databases. The author reviewed three representative

[107] On 1 January 2015, ILTEC and the *Centro de Estudos de Linguística Geral e Aplicada* (CELGA) ('Centre for Studies of General and Applied Linguistics') merged, forming the Research and Development Unit of the University of Coimbra named CELGA-ILTEC. http://celga.iltec.pt/pt/news.html.

[108] (…) "*outros projectos surgiram que têm como objectivo a construção de gramáticas e dicionários electrónicos e de interfaces em língua natural, a criação de bases de dados terminológicas e de corpora informatizados, de correctores e analisadores morfológicos, de instrumentos de tecnologia educativa. O Português encontra-se aí em perfeita igualdade com as línguas maioritárias europeias*".

projects promoted by the European Union, namely, MULTILEX (A Multilingual Standardized Lexicon for the European Community Languages), ACQUILEX (The Acquisition of lexical knowledge for Natural Language Processing systems), and GENELEX (acronym for GENeric LEXicon).

She also reported on efforts that have been made for standardization of work on language technology resources in order to 'allow the reusability of the databases, the sharing of information, the economy of resources and the faster development of this discipline'[109] (Correia, 1994, p. 11; my translation). This remark is particularly telling of the context within which these first language technology enterprises originated. Thus, Correia (1994) presented the lexicographic branch of the recently created (February 1993) EAGLES[110] (Expert Advisory Groups on Linguistic Engineering Standards), developed within the Linguistic Research and Engineering Programme, with funding from the European Commission. Correia and Guerreiro (1995, pp. 65-66) succinctly reviewed previous projects setting standards for digital data reusability: the Text Encoding Initiative (TEI), ET-7 (Standards for Reusable Lexical and Terminological Resources), and NERC (Network of European Reference Corpora).

Among many projects using digital lexical databases developed in the 1990s in Portugal, Correia and Guerreiro (1995) highlighted the *Dicionário de Termos Linguísticos*[111] and *Dicionário de Termos Informáticos*[112] ('Dictionary of Linguistics Terms'; 'Dictionary of Computational Terms') (ILTEC); *Observatório do Português Contemporâneo* ('Observatory of Contemporary Portuguese') (University Nova de Lisboa); Lince spellchecker (ILTEC); and the *Corpus de Referência do Português Contemporâneo* (CRPC) ('Corpus of Reference of Contemporary Portuguese'), by Professor João Malaca Casteleiro, coordinated by Maria Fernanda Bacelar do

[109] "(…) *garantir a reutilizabilidade das bases, a partilha de informação, a economia dos meios e o mais rápido desenvolvimento desta disciplina*".

[110] In 1996, John Sinclair shared with participants of the *XI Encontro da Associação Portuguesa de Linguística* ('XI Meeting of the Portuguese Association of Linguistics') the recommendations for Classification of Text Types adopted by EAGLES, which should be accompanied by the Classification of Corpus Types (Sinclair, 1994), as suggested by the author (1996, p. 39).

[111] First published in a printed version in two volumes. An updated online version is freely available at http://www.portaldalinguaportuguesa.org/?action=dtlinginfo.

[112] ILTEC. (1993). *Dicionário de Termos Informáticos*. Ed. Cosmos.

Nascimento and compiled by a team of researchers at the *Centro de Linguística da Universidade de Lisboa*[113] ('Centre of Linguistics of the University of Lisbon').

It can be said that the project for the compilation of the CRPC represented an important step forward in the history of Portuguese corpora. With the challenging aim to cover all existing Portuguese language varieties (not only from CPLP but also Goa and Macau), in addition to diverse text types (books, newspapers, magazines, interviews, etc.) and modes (oral and written), the enterprise started in 1988, and by 1995 the corpus had reached the size of 30 million words[114]. The corpus was POS-tagged with an adapted version of EAGLES tagset incorporated in the PALAVROSO POS-tagger, developed by INESC (Portugal) (see Nascimento & Gonçalves, 1996). As Nascimento and Gonçalves (1996) indicate, the construction of CRPC has also greatly benefitted from the corpus compilers' unique experience of working in European-funded projects, for instance, PAROLE (Preparatory Action for Linguistic Resources Organization for Language Engineering). Moreover, CRPC was used as the source for examples extraction in the *Dicionário da Língua Portuguesa Contemporânea* (2001).

Without a doubt, dramatic progress in computational linguistics made in the 1990s led to a significant expansion of corpus building. Along with CRPC, many other corpora have been compiled, whereas digitization of extant manually constructed ones resulted in the unparalleled wide availability of previously restricted-use sources. For an elucidative account of existing corpora of Portuguese hosted by Portuguese institutions in 1996, see Nascimento, Rodrigues and Gonçalves (1996, pp. 423–447). An attempt to present thorough information about corpora created in Brazil was made by Castilho, Silva and Lucchesi (1996, pp. 113-128).

Given what has been shown, the participation of Portugal in the projects for the creation of language technology resources promoted by the European Commission in the 1990s can be considered to have actively fostered the continuation of investigation and development of Portuguese language processing methods and tools production. Furthermore, this multinational work experience has equipped Portuguese researchers

[113] http://www.clul.ulisboa.pt/en/

[114] As of 17th of April 2017, the corpus has 309,812,943 words (cf. http://alfclul.clul.ul.pt/CQPnet/crpcnetfg/index.php?thisQ=corpusMetadata&uT=y). More details on this corpus in the next section.

with the most advanced knowledge, thus solidly foregrounding further development of this lexicographic computational area in Portugal.

In the same period in Brazil, there was also a significant increase in the development of language technology resources, however, with a much lower magnitude. A milestone in the history of Brazilian computational linguistics was the foundation of the pioneering Interinstitutional Centre for Computational Linguistics (NILC) in 1993 at the *Universidade de São Paulo* (USP) at the São Carlos campus.[115] A key product developed in this centre was ReGra, which is the grammar checker for Brazilian Portuguese integrated into Microsoft Word and the word processor REDATOR by Itautec-Philco.

The beginning of the building of the Corpus NILC was another crucial moment in the history of computational resources for supporting language investigation in Brazilian Portuguese. This corpus was initially created as a testbed for ReGra, thus comprising three categories of texts: with correction, partial correction and no correction. The great success of this corpus among the Natural Language Processing community in Brazil, given that it was the only large corpus of Brazilian Portuguese available, led the compilers to propose its expansion, making Corpus NILC part of the future Corpus Lácio-Web.

Other large and carefully designed corpora that began to be compiled at around the same time (in the 1990s) was the *Corpus de Araraquara*, initially containing some 30 million words. According to Neves (2011), the creation of this corpus was a joint initiative of Professor Francisco Borba and Professor Maria Helena de Moura Neves, at the *Universidade Estadual Paulista "Júlio de Mesquita Filho"* (UNESP) (also known as 'State University of São Paulo') – Araraquara, in order to serve as the basis for the creation of two major pedagogical resources still lacking (at that time) in the context of Brazilian Portuguese: both a dictionary and a usage grammar of Portuguese. The former was published in 2002 under the coordination of Prof. Borba, while the latter came out in 2000, authored by Prof. Moura Neves.

[115] Later on, other institutions have become part of the centre. According to the initial page information of their website, "it now includes computer scientists, linguists and research fellows from several universities and research centres, such as the Federal University of São Carlos (UFSCar), State University of São Paulo (UNESP), and State University of Maringá (UEM), among others." Available at http://www.nilc.icmc.usp.br/nilc/index.php.

As indicated in the initial pages of the *Dicionário de Usos do Português do Brasil*, a 70 million-word corpus[116] was the source for extraction of the approximately 62,000 headword list (Borba, 2002). It is also declared that the detailed syntactic-semantic information in the entries reflects the real use of Portuguese in Brazil. Particularly significant is Biderman's evaluation of this lexicographically-modern work. The author has argued that the metalanguage employed demands full linguistic knowledge from the user and makes the entry overloaded with grammatical information, thus hindering its reading and comprehension (Biderman, 2003b, p. 62).[117]

Construction of large and varied electronic corpora of Portuguese experienced steep growth in the 2000s due to the advantages brought forth by the Internet, like ease of access to texts sources, relaxation of copyright issues and development of user-friendly tools for web crawling. Nowadays, a great number of different corpora are freely available for interrogation and download. Furthermore, additional computational tools for corpus analysis have been produced and most of them are free for online use or download.

Since it is not feasible to review all the digital resources currently available for the Portuguese language here, an indication of some of the most comprehensive repositories of corpora, resources and tools is given below.[118] See Appendix B for print screens of their home pages, which provides an illustrative overview of their features.

- **CEPRIL[119]:** stands for *Centro de Pesquisas, Recursos e Informação em Linguagem* ('Centre for Language Research, Resources and Information'), at the *Pontifícia Universidade Católica de São Paulo* (PUC-SP). It is coordinated by Professor Tony Berber Sardinha, who introduced corpus linguistics in Brazil. One of the major achievements of this group was the compilation of the 1 billion token *Corpus Brasileiro*,

[116] This is the *Corpus de Araraquara*. In 2011, it had been enlarged to over 200 million words (Neves, 2011, p.35). Unfortunately, the *Corpus de Araraquara* is not publicly available.

[117] It seems apparent that it is difficult, even for highly experienced and respected lexicographers such as Borba and Vilela (see the case of the *Dicionário Básico de Português*, mentioned earlier), to strike a balance between innovation in a dictionary's layout and the type of information provided (here, syntactic patterns) and user-friendliness. This lesson should be kept for later when planning DOPU design.

[118] However, it should be stressed that most universities and research centres working on Portuguese showcase their work on websites, with either free (registered or unregistered) access to online functionalities or full download.

[119] http://www.pucsp.br/lael/cepril/cepril-info.php

- **CLUL**[120]: as mentioned above, the Centre of Linguistics of the University of Lisbon, which was founded in 1932, has a rich website with access to all electronic and digitized products developed within the various research lines. The availability of important resources for the treatment of spoken Portuguese and a corpus of old texts, which allows detailed analysis of texts written or translated into Portuguese until the year 1525, should be noted.
- **Linguateca**[121]: created in 1998 with Diana Santos as its mentor, this website is undoubtedly the most varied and largest source of language technology support regarding the Portuguese language. As a computational linguist with strong beliefs in the open-source ideal and free accessibility to everyone with an interest in the subject, this site offers resources, corpora and tools either developed by members of Linguateca or external researchers, together with several links to very instructive additional information. For an elucidative account of Linguateca's infrastructure, see Santos (2011) and Santos (2015).
- **LX-Centre**[122]**:** The Language Resources and Technology for Portuguese Centre is a web site offering access to a series of services (for online use of tools), tools (for download), applications (for online use of programs like translator and summariser), and a variety of datasets. This Centre belongs to the NLX-Group, which is the Natural Language and Speech Group, led by António Branco in the Department of Informatics of the University of Lisbon, Faculty of Sciences.
- **NILC**[123]: the website of the centre previously mentioned provides free access to use and/or download all tools and resources developed by their team. These comprise writing assistance and text simplification and evaluation; corpora; lexicons and semantics; syntax; semantics and discourse; summarization; preprocessing tools and part-of-speech tagging; machine translation and speech.

Given what has been shown so far, it could be said that the Portuguese language has a quite long and productive history with regard to digital resources. Nevertheless, it

[120] http://www.clul.ulisboa.pt/en/resources-en

[121] http://www.linguateca.pt/

[122] http://lxcenter.di.fc.ul.pt/home/en/index.html

[123] http://www.corpuslg.org/tools/

has been argued that language technology support and linguistic research still have to be much more advanced to achieve "consolidation of the Portuguese language as a language of international communication with global projection" (Branco et al., 2012, p. 42) since "in the case of Portuguese, language technology support has been steadily improving but it requires a strategic boost to reach decisive level of sustained development" (Branco et al., 2012, p. 71).

Although Branco and colleagues' claim was made five years ago, it seems possible to affirm that the desire for further, qualified Portuguese human language technology development still holds true. This PhD thesis testifies to that. In order to apply state-of-the-art methodology for dictionary-making, a series of other resources had to be created, as will be shown in part II. Among them, I highlight here the key element of this project, which is fundamental for building a corpus-driven dictionary: the corpus.

3.3.1 Corpora of Portuguese with academic texts

DOPU's target users are students in higher education, attending courses in different areas of knowledge, whose language of instruction is (Brazilian or European) Portuguese and thus need to read and write academic texts in Portuguese. As a corpus-driven dictionary it must portray the linguistic information that is based on texts that reflect the way language is used by expert writers from Brazil and Portugal in academic settings in different areas of knowledge. Hence, the corpus needed for making DOPU must be: composed of academic written texts portraying exemplary language; balanced in terms of Portuguese varieties: 50% of Brazilian Portuguese, 50% of European Portuguese; discipline-varied, i.e. covering different academic areas; synchronic; and large in size.

The first step in the process of conceptualization of DOPU was to examine existing Portuguese corpora containing academic texts and determine their suitability for my research. Out of many corpora of Portuguese in existence (as mentioned in the previous subsection), which cover different language varieties, registers, and genres, only a few comprise academic texts. As Table 3.4[124] shows, although existing corpora

[124] This table was published in Kuhn and Kosem (2016).

of Portuguese do contain academic texts, none of them gathers all the characteristics mentioned above. Consequently, a decision was made to compile a new corpus of academic texts, which I named *Corpus de Português Escrito em Periódicos*-CoPEP ('Corpus of Portuguese from Academic Journals') (see Chapter 6 for a full account of the compilation process and presentation of the corpus).

Table 3.4 Suitability analysis of Portuguese corpora with academic texts

Corpus and author(s)	Size	Characteristics	Reasons for not suiting my purposes
Portuguese Web 2011 (ptTenTen, Palavras parsed) Authors: The Sketch Engine team	2,757,635,105 words[125]	Texts from sites of academic/scientific nature (universities, journals, governmental, thesis repositories, etc.) Parsed by PALAVRAS dependency parser (Bick 2000).	Crucial metadata such as source (type of publication: journal, book, thesis, etc.), year of publication and area of knowledge are not available. No possibility to measure quality of writing and corpus composition.
Portuguese Web 2011 (ptTenTen, Freeling v3) Authors: The Sketch Engine team	3,900,501,097 words	Texts from sites with academic/scientific nature (universities, journals, governmental, thesis repositories, etc.) Tagged by Freeling 3.0 (Padró & Stanilovsky 2012)	Crucial metadata such as source (type of publication), year of publication, area of knowledge and language variety are not available. Country of the website is made equivalent to language variety, which is not an accurate approach for determining such relevant information. No possibility to measure quality of writing and corpus composition.
Corpus Araneum Portugallicum Maius (Portuguese, 15.05) 1,20 G as a language resource Author: Vladimír Benko	862,134,902 words	Texts from sites of academic/scientific nature (universities, journals, governmental, thesis repositories, etc.). To be used for contrastive linguistics and bilingual lexicographic projects.	Crucial metadata such as source (type of publication: journal, book, thesis, etc.), year of publication, area of knowledge and language variety are not available. No possibility to measure quality of writing and corpus composition.
Corpus Brasileiro ('the Brazilian Corpus') Author: Tony Berber Sardinha (coordinator)	1,133,416,757 tokens	General corpus of Brazilian Portuguese. Academic subcorpus contains 258,585,002 tokens from articles, 310,972,387 tokens from theses and dissertations, and 6,947,244 tokens from annals.	Crucial metadata such as year of publication and area of knowledge are not available. No information on quality of texts comprising the academic subcorpus. Only Brazilian Portuguese.
Corpus do Português (Genre/historical version) (the Corpus of Portuguese) Authors: Mark Davies and Michael Ferreira	45 million words	Texts of the 1300s to the 1900s. The texts from the 1900s make up 20 million words, with balance between academic, fiction, spoken and newspaper genres. Its academic subcorpus consists of 3,087,052 words from Portugal and 2,816,802 from Brazil.	Academic subcorpus is composed of entries retrieved from Brazilian and Portuguese online encyclopaedias.
CPBA – *Corpus do Português Brasileiro Acadêmico* ('the Academic Brazilian Portuguese Corpus') Authors: The research group UPLA, coordinated by Cristina Becker Lopes Perna, at PUCRS (Brazil)	22,777,993 tokens (Peixoto, 2015, p. 44)	Books and journals from six different areas of knowledge provided by eight Brazilian universities comprising written productions of professors and (undergraduate and graduate) students.	Not publicly available. Only Brazilian Portuguese.
CRPC - Corpus de Referência do Português Contemporâneo ('Reference Corpus of Contemporary Portuguese') Authors: developed at the *Centro de Linguística da Universidade de Lisboa* (CLUL).	311 million words (spoken+written) approx. 310 million words of written texts	General language corpus. European Portuguese and other varieties (Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, S. Tome and Principe, Goa, Macao and East-Timor). Comprising different text types, including scientific. Texts from the second half of the 19th century to 2008.	Metadata not consistently available.

[125] In this table, words, tokens, or both, are used when providing information on corpus size, depending on the information that is available.

3.4 Concluding remarks

This section has reviewed the relationship between corpora and lexicography. It has shown that the advent of computational technology has had a huge impact in both language studies and lexicography, resulting in new fields such as corpus linguistics and e-lexicography.

With regards to the Portuguese language, it has been shown that, although Brazil and Portugal have taken some important steps towards the use of computational technology in such areas, we still lag quite behind other languages. Among some of the gaps that need filling, two have been spotted that play key roles in the design of DOPU: lack of a well-designed and carefully planned corpus of academic Portuguese and no previous adoption of the semi-automated approach to dictionary-making. In consequence, for my research project to be realised I needed to compile a new corpus and develop all the required tools and resources for application of the automatic approach. Part II of this thesis describes the process of resources and tools development specifically for DOPU. But before that, a detailed design plan has to be set out. The next chapter accounts for that.

Chapter 4 Planning the *Dicionário de português para estudantes universitários* (DOPU) – Dictionary of Portuguese for university students

This chapter sets out to present the process of planning DOPU. It will exhibit the factors that were taken into account, beginning with a characterization of the envisaged dictionary and its target users, then moving on to the definitions of crucial points of decision regarding content and structure. The purpose of this plan is to provide a framework that would enable hands-on linguistic and lexicographic work to be carried out in a sound and consistent manner, as will be shown in Chapters 5 to 11.

As a means of organization, four stages were defined: pre-compilation definitions (4.1), lexicographic evidence acquisition (4.2), candidate headword list building (4.3) and entry compilation (4.4).[126] This allowed me to have a clear overview of each stage, how they connect to each other and what the requirements are for the concatenation of these parts into one complete, final design proposal.

For each stage, I envisaged which goal(s) should be reached, then listed key factors that needed to be accounted for in order to accomplish those goals, and, bearing in mind the final objective, that is, to propose a design of DOPU, I made decisions of how to go about the development of the design. Table 4.1 below is a graphic of this preparation phase.

Table 4.1 Development of DOPU's design

Phase of lexicographic work	Decisions involving:	Objective of this PhD research
Pre-compilation	Characteristics of the dictionary; Target users	*Definition of a user profile*
Lexicographic evidence acquisition	Method of data acquisition	*Experiment of automatic extraction of data from the corpus and import into dictionary writing system*
Candidate headword list building	Macrostructure: • headword list • lexical entries	*Provision of principles for candidate headword list building*
Entry compilation	Microstructure: • microstructural components Entry compilation: • primary source • secondary sources • additional content	*Provision of principles for entry-writing*
		Proposal of a design of DOPU

[126] I loosely followed Atkins and Rundell (2008), Hartmann (2001), Kiefer and van Sterkenburg (2003) and Svénsen (2009).

If read horizontally, from left to right, the table shows each stage of the preparation phase of the lexicographical process (column 1), factors about which decisions have to be taken (column 2), and the objective that is achieved by making those decisions (column 3). If column 3 is read vertically, from the top to the bottom, the achievement of all those objectives makes up the design proposal of DOPU.

It should be noted that the design does not encompass those aspects of the dictionary with regard to the final online tool. So, for instance, hyperlinking, user interface interaction, dictionary layout, among others, will not be developed in the design. Rather, only suggestions of useful features will be briefly touched upon, which is presented here in section 4.5. They are mentioned in this chapter, at any rate, due to their role in the conceptualization of DOPU. All things together, a great deal of decision-making depends on the envisaged final product as a whole, especially the technology required for actual implementation.

Before moving on, it is crucial to highlight a key feature of DOPU: it is corpus-driven. That means that the corpus is used not only for writing entries but also for directing decisions for all other dictionary contents. The choice for a corpus-driven approach for this lexicographic project derives from the attested fundamental role of corpora in dictionary-making:

> On the *macrostructural* level corpora provide crucial information for the creation of the lemma-sign list of a dictionary, and on the *microstructural* level corpora enable lexicographers to tremendously enhance the accuracy of the dictionary articles themselves. (de Schryver & Prinsloo, 2000b, p. 292)

4.1 Pre-compilation definitions

It is widely known that the first step in a project making a new dictionary concerns the definition of the type of dictionary and the user profile,[127] that is, who the users are and what their needs are (cf. Atkins and Rundell, 2008;Gouws, 2010; Grundy

[127] Although the appeal to consider users' needs in the process of dictionary-making is said to have begun in the 1960s, when Householder affirmed that "Dictionaries should be designed with a special set of users in mind and for their specific needs" (Householder, 1962, p. 279) (as cited in Welker, 2010), Lew affirmed that only recently have users become a crucial element in dictionary design (Lew, 2015a).

& Rawlison, 2015; Klosa, 2013; Nesi, 2012, among others). As Lew (2015a, p. 2) explains:

> specifying the (foreseen) characteristics of the target user at the planning stage helps in the design of dictionaries which do not yet exist, by equipping them with the lexicographic data that are likely to be expected and used.

Thus, the design of DOPU began with a thorough description of the type of dictionary that I envisaged, followed by a detailed characterization of the target user. It should be stressed that, although these two phases are presented separately here, in cognitive terms, they were simultaneous. All the following decisions in each different stage were taken in order to comply with the needs of a particular target group using a specific dictionary for a determined purpose.

4.1.1 Type of dictionary

The *Dicionário de português para estudantes universitários* (DOPU) has been conceptualised to be an online corpus-driven dictionary of Portuguese for university students. Following Atkins and Rundell's (2008, pp. 24-25) simple questions-and-answers system for defining the properties of dictionaries,[128] this is the characterization for DOPU:

[128] These categories are also tackled in other lexicographic literature concerning dictionary typology, e.g. Landau, 2001; Svénsen, 2009; Zgusta, 1971, among others.

Table 4.2 DOPU characterization

The dictionary language?	Brazilian Portuguese (BP) European Portuguese (EP) Monolingual
The dictionary's coverage?	Written academic language Variety of areas of knowledge
The dictionary size?	Number of entries to be corpus-driven determined
The dictionary's medium?	Online
The dictionary's organization?	Word to meaning (semasiological)
The user's language?	Speakers of BP as a mother tongue Speakers of BP as an additional language Speakers of EP as a mother tongue Speakers of EP as an additional language
The user's skills?	Undergraduate university students Graduate university students Inexperienced users of academic Portuguese PAL speakers: advanced level
What is this dictionary for?	Production and reception

4.1.2 User Profile

First and foremost, as a corpus-driven dictionary designed from scratch, determination of texts to create the corpus depends on the kind of information that should be displayed in the dictionary, which in turn is defined by the needs of the users.

Consequently, it can be said that the definition of the user profile is the grounding pillar when one attempts to build the best dictionary as possible for the target users as it is connected to all design phases: corpus creation, lexicographic evidence acquisition and macro- and microstructural aspects.

According to Atkins and Rundell (2008, pp. 28-30), description of the user profile concerns envisaging the characteristics of the type of user that will use the dictionary and what the purposes of its use will be.

An important contribution that helps to delineate a user profile derives from user research (Atkins & Rundell, 2008, pp. 30-32), which is academic experimental research focusing on dictionary use. User research is a niche of studies within *metalexicography*, as Hausmann, Reichmann, Wiegand, and Zgusta (1989) announced in the preface of the

Wörterbücher, Dictionaries, Dictionnaires. An International Encyclopedia of Lexicography:

> [Metalexicography] is the scientific discipline which studies dictionaries, their forms, structures, and uses; their criticism and history, their position in society; the methodology and procedures of their compilation, and their underlying theoretical stances. (Hausmann et al., 1989, p. xvii)

Interesting findings from empirical research like users' preferences (dictionary format, entry layout, definition), habits (look-up strategies), attitudes (opinions on dictionaries and dictionaries' elements, suggestions) and context of use (decoding, encoding, work, school, leisure) can be very useful when making lexicographic decisions of various orders.

For the definition of DOPU's user profile, first the types of users and DOPU's envisaged functions – in terms of what is expected to be users' demands – will be presented. Then, the contribution of user research will be only briefly discussed since the lack of studies on users of Portuguese dictionaries[129] led to the appropriation of some findings of existing experiments for other languages that could be applicable to my project.

Overall, as Atkins and Rundell's (2008, p. 28) cleverly put it: "Know your users: that way, the dictionary will give them what they need".

4.1.2.1 Types of user

As is shown in the name DOPU, this dictionary is for university students, both undergraduates and graduates. In general terms, the difference between the two groups lies in the level of familiarization with academic language: undergraduates are novice users of such language use, whereas graduates are more experienced users.

[129] This gap will be soon bridged. The European Survey on Dictionary Use is being conducted at the time of writing (May 2017), contemplating 29 countries, 26 languages and involving 58 researchers. I am on of the representatives of the Portuguese language; thus, answers will be collected in Brazil and Portugal and data will be statistically analysed together with other languages. Attention should be drawn to the fact that it will be the first time in the history of metalexicography of the Portuguese language that such fundamental information on users' attitudes, preferences, habits, and context of use will be obtained. For more information on the survey and participating researchers, see http://www.elexicography.eu/events/european-survey-on-dictionary-use/

At this point, it should be stressed that DOPU is targeted at university students who use Portuguese as a medium of instruction, independently of their mother tongue. The reason for this choice is my understanding that academic language is a register of language that requires learning, as shown in detail in Chapter 2. Speakers of Portuguese as a mother tongue have not experienced academic language in school, nor have speakers of Portuguese as an additional language, thus both need an auxiliary lexicographical tool for higher education literacy practices.

Given that students are not familiar with academic language, that is, with the way known words are used in particular ways, in different combinations and with certain meanings, the language covered in DOPU is the one used in academic contexts. Alonso, Millon, and Williams (2011, p. 12) provide a compiling argument in favour of my point of view:

> In reality, specialised communication is not just about technical words. In most cases, scientists already know the definition of the technical word, but look up the 'specialised' meaning of a general word in the dictionary, for getting information of the behaviour of the word in a domain-specific context.

It is expected that speakers of PAL have an upper-intermediate or advanced proficiency level, which means a good command of the language to follow the courses and engage in different academic activities.

Evidently, mastery of Portuguese might vary. For that reason, the decision was taken to accommodate DOPU's design to speakers of PAL who would benefit from more pedagogically-enriched features. While it can be argued that speakers of a language as a mother tongue would not need these more explanatory elements and that they could hinder the look-up, it is also true that one of the unquestionable advantages of electronic dictionaries is customizability, meaning that some features can be shown only when chosen, thus catering for different users' need in only one place (cf. Lew, 2015a), as will be further addressed in section 4.5 below.

What is more, Nesi (2012) shows that such a clear-cut differentiation between dictionary for learners and for speakers of a language as a mother tongue has been fading, "because many of the innovative features that were first introduced to help language learners, such as usage notes, corpus-based examples, and writing guide

sections, are gradually being included in mainstream L1 dictionaries edition" (p. 2). She also refers to her study (Nesi & Haill, 2002), which indicated use of monolingual dictionaries of general language instead of learners' dictionary by some learners.

As learners of academic language, it is predicted that users will use DOPU for comprehension and production purposes. Thus, while monolingual dictionaries of general Portuguese tend to focus on receptive functions, providing definitions that explain the meanings of words and usually using made-up examples or literary citations, DOPU will provide explanatory paraphrases with simpler language and production-oriented features, thus attending to both functions. Some of the well-known features in dictionaries for text productions are the presentation of collocations, information on frequency of occurrence, elicitation of use patterns and syntactic structure, and usage notes.

Users of DOPU will speak Brazilian Portuguese and European Portuguese. While in the past unification of both varieties in one lexicographical tool would be frowned upon due to widely known reasons concerning important differences that cannot be simply overlooked, once more the advent of technology subverts old traditions. Nowadays, one dictionary can hold information that is specific to different uses and users. In terms of language variety, there are two possibilities to customise this aspect, both widely adopted in other lexical resources: users' choice of language variety before starting to look up or automatic assignment of language variety due to IP address identification. Once language variety is determined, only information stored in the database with regard to the chosen variety will be displayed (cf. Atkins, 1996, as cited in Varantola, 2002; Lew & de Schryver, 2014).

Finally, users will study different subjects. In this way, it would be best to adapt the look up result to their specific needs. Although this feature implementation, as with the previous one on language variety, depends highly on the work of computational linguists and web designers– thus, beyond my abilities – it is important to predict such a possibility. For now, a possible alternative is to present examples of uses of the headword in different areas of knowledge.

The dictionary user characterization presented here defines the user profile of DOPU, which set the ground for corpus compilation and macrostructural and

microstructural decisions. However, as Atkins and Rundell (2008, pp. 30-32) suggest, empirical user research can be of great help as well. Thus, in the next section, I will refer to some experiments carried out with other languages and select some interesting findings for the design of DOPU.

4.1.2.2 Research on dictionary users

With Barnhart's pioneering questionnaire survey on dictionary use in 1955 (Welker, 2010, p.12), it can be said that the user perspective started to gain new ground. However, it should be noted that the development of this type of research was rather slow. Welker (2010) showed that only six empirical studies on dictionary use were undertaken before 1980. Wiegand's interest in determining a "sociology of dictionary use" in the mid-1970's (Lew, 2004; Welker, 2010) played an important role in the change of this scenario, as it "cleared the way for modern empirical research on dictionary use" (Lew, 2004, p. 35).

As to the methodology used for research on dictionary use, researchers (Duran, 2008; Dziemianko & Lew, 2006; Koplenig, 2014; Lew, 2004, 2015b; Welker, 2010) referred to the use of questionnaires, interviews, observation, (written and oral) protocols, tests and experiments, and log-files. Quite recently, due to the medium-change from paper to digital and all the still-unknown implications for dictionary use, research employing eye-tracking technology and usability paradigms has begun to be carried out (cf. Lew 2015a, 2015b).

For an enlightening outline of a great number of studies on dictionary use, Welker's *Dictionary use: A general survey of empirical studies* (2010) is a monumental book which aims to present a comprehensive overview of all empirical studies on dictionary use ever done. The author provides summaries and bibliographic annotations of 320 empirical studies, encompassing not only dictionary use research in the context of English language learning (which has been the focus of the majority of the studies), but also dictionaries of other languages and their users. He divides his reports according to the type of inquiry undertaken: surveys; studies of actual dictionary use; studies of the effects of dictionary use; studies of specific dictionary features and of specific dictionaries; research on the use of electronic dictionaries; and on the teaching of dictionary use.

It is significant to point out that, out of the 320 scrutinised studies in Welker (2010), only 30 (9.4%) refer to dictionary use research that involves the Portuguese language: an investigation of the use of Portuguese school dictionaries by Brazilian children and elementary school teachers; 27 concern the use of bilingual dictionaries, with foreign language the focus of the study; and one is a case study reported by a North-American linguist who used a bilingual (English-Portuguese/Portuguese-English) dictionary as an auxiliary tool in the process of learning Brazilian Portuguese. Bibliographic reviews for more recent research adopting an empirical approach to the use of Portuguese dictionaries revealed the existence of a few topic-specific studies, confirming that this subject is underrepresented in the metalexicography of the Portuguese language[130].

Lack of representative empirical research on Portuguese dictionary use, let alone one that is university-student oriented, led me to access studies on other languages and adopt some findings that contribute to making decisions for DOPU design.

An attempt has been made to collect outcomes from more recent studies, since the bulk of the research in the area is quite outdated, concerning mostly the use of paper dictionaries and the early consultations of electronic dictionaries. As Lew and de Schryver (2014) observed, nowadays, with the digital world being part of one's life, "the status of the dictionary is changing, and so are patterns of user behaviour" (p.341). Table 4.3 presents the findings and their implication for DOPU's design:

[130] Some of the studies found were: Bolzan (2012); Lugli and Silva (2012); Ventura (2014); Costa, Rebouças, and Pontes (2014).

Table 4.3 Dictionary user's research findings

Findings	Implications for DOPU's design
Meaning is the top reason for dictionary consultation (Lew,2010b)	Choice of definitions styles Contemplation of use of defining vocabulary
Pictorial illustrations are helpful additions to the definitions, contributing to comprehension of concrete nouns (Gumkowska, 2008) and vocabulary acquisition (Nesi, 1998) (as cited in Lew, 2010a)	Use of images in definitions of concrete nouns
Use of animations in definitions has shown to be counterproductive (Lew & Doroszewska, 2009)	Save the use of animations for cases where definitions would get very complicated otherwise.
50% of 620 university students reported looking at only the first sense of the word (Kosem, 2010)	Academic sense of the word should be first Customization of sense order according to area of knowledge
The microstructural features used the most by university students are definitions, synonyms, spelling, and examples; while frequent phrases, usage and grammar, collocates and pronunciation were rarely used by native speakers but quite frequently used by non-native speakers[131]. (Kosem, 2010)	Customization of microstructural features display according to status of Portuguese for the user
The activity of writing academic work was ranked first with regards to university students' opinions on the importance of dictionary use, followed by reading academic books and journals (Kosem, 2010)	Display of helpful information for productive purposes, like syntactic structure, pattern of use, usage, synonyms, collocations, examples
As to preferred display of access to entry components, namely, grammar, paraphrase, typical contexts, and sense relations, tabbed interface – the one that resembles a web-browser interface - was ranked highest among users of online dictionaries (Koplenig & Müller-Spitzer, 2014)	Web designer should be instructed to create DOPU's layout to display entry information in tabs
Display of three examples rather than just one is considerably more helpful (Frankenberg-Garcia 2012, 2014)	Set script of automatic extraction of data to select the top three examples as ranked according to GDEX
Users adopt typical WEB search engine strategies for looking up in electronic dictionaries, typing in keywords, phrases, whole sentences, and even full paragraphs[132] (de Schryver, 2006)	Create a search mechanism that maps the typed word form to the lemma (cf. Seretan & Wehrli, 2013) Use "did you mean...?" function for spelling errors Use auto-completion of words' initials in form of a drop-down list for users to choose from
Users have declared to expect online dictionaries to contain reliable and up-to-date content and clarity of presentation (Müller-Spitzer, 2014)	Include the dictionary in the renowned *Portal da Língua Portuguesa*[133], which is supported by FCT[134] (the Portuguese national funding agency for science, research and technology)

Overall, an elaborated characterization of DOPU, detailed description of target users, and attentive collection of relevant user research findings produced a clear, well-defined picture of the users' profile of DOPU.

131 Kosem alerts that the reason for the very rare use by native speakers of these four last features might be due to their low prominence in general dictionaries or simply inexistence, and not necessarily a lack of interest (Kosem, 2010, p. 168).

132 Results obtained from log-files.

133 http://www.portaldalinguaportuguesa.org

134 http://www.fct.pt/

4.2 Lexicographic evidence acquisition

Atkins and Rundell have affirmed that "objective evidence of language in use is a fundamental prerequisite for a reliable dictionary" (2008, p. 53). In order to access real instances of language in use, lexicographers need to work with corpora. As shown in Chapter 3, since the advent of electronic corpora, corpus-based/driven lexicography has become the norm, and DOPU could not depart from that principle.

As concerns the design of DOPU, employment of a corpus-driven approach means that the corpus is used to inform every aspect of the dictionary: from candidate headword list extraction to syntactic behaviour description, from meaning disambiguation to examples selection. In other words, macrostructural and microstructural decisions in this PhD research were taken based on what the corpus provided.

A crucial decision to take was thus: what approach should be used for acquisition of knowledge from the corpus? Among the most advanced corpus-based methods for dictionary-making, it has been shown that the semi-automated approach stands out. By taking the relationship between man and machine a step further towards automation, this state-of-the-art method has contributed to a considerable enhancement of the lexicographical process.

In this vein, it was apparent that DOPU should take the greatest advantage of recent developments in lexicographical work, especially as DOPU is a digitally-born dictionary created from scratch. It was then my decision to adopt the semi-automated approach for conceptualization and creation of the design of DOPU.

4.2.1 The semi-automated approach to dictionary-making

Recently, a ground-breaking method has been developed in the context of the Slovene language in which data is automatically extracted from the corpus and imported into the dictionary writing system (see Chapter 3). Thus, the point of departure of lexicographic work is no longer concordance lines or word sketches, but rather pre-populated entries in the DWS. In consequence, the time-consuming tasks of manual selection of collocates and examples for each grammatical relation, together with the

tedious routine of copying the data from the corpus tool and pasting them into the dictionary writing system, were actually taken over by the computer.

It has been shown that this shift from humans to computer has considerably streamlined lexicographic work, allowing more time for lexicographers to spend on analytical and editorial activities (Gantar et al., pp. 218-219), such as sense identification, definition writing, and collocates sorting. This method has been called the semi-automated approach to dictionary-making (Gantar et al., 2016).

According to Gantar et al. (2016, pp. 220-221), the procedure of automatic data extraction is language-independent. However, the following requirements have to be met:

a) Software: corpus tool and dictionary writing system
b) Corpus with part-of-speech (POS) annotation
c) Sketch grammar
d) GDEX configurations
e) Definition of extraction procedure

The basic data extracted from the corpus are grammatical relations, collocates and examples. Additional information can be acquired, as will be shown in section 4.2.1.5 below.

4.2.1.1 Software

Although this method can be used with any software, in this thesis I followed very closely the procedure implemented by the Slovene team, which used the Sketch Engine (Kilgarriff et al., 2004) corpus tool and iLex (Erlandsen, 2010) dictionary writing system. This decision is in line with de Schryver and Prinsloo's (2000a, p. 312) wise remark, “a corpus without advanced corpus query tools is of no use.”

At this point, it should be mentioned that the Sketch Engine and iLex are paid programs. While the former works with a subscription system, which is valid for a

determined period of time and allows unlimited renewals, iLex sells a one-time licence. In the case of this PhD research, I was given access[135] to the Sketch Engine and iLex.

The Sketch Engine is considered the most sophisticated corpus tool currently available. Its hallmark is the word sketch function, which provides a summary of the grammar behaviour of a headword. According to its creator, Adam Kilgarriff,

> the word sketch can be seen as a draft dictionary entry. The system has worked its way through the corpus to find all the recurring patterns for the word and has organised them, ready for the lexicographer to edit, elucidate, and publish. (Kilgarriff et al., 2014, p. 10)

Figure 4-1 below shows a partial word sketch for the noun *análise* ('analysis') in the grammatical relations noun+ (participial) adjective, verb+object (noun) and noun+and/or+noun.

análise_N mod por Adj-Part			V obj análise_N			e_ou		
	14,116	0.32		3,608	0.08		1,441	0.03
fatorial ✚	1,058	11.06	realizar ✚	279	9.61	discussão ✚	156	11.29
análise fatorial			realizar uma análise			análise e discussão		
estatístico ✚	1,203	10.96	efetuar	96	9.27	interpretação ✚	148	11.19
análise estatística			efetuar uma análise			análise e interpretação dos		
comparativo ✚	514	10.04	permitir ✚	185	8.90	comparação	42	9.69
análise comparativa			fazer ✚	346	8.76	reflexão	38	9.31
descritivo ✚	415	9.70	fazer uma análise			análise e reflexão		
análise descritiva			centrar	44	8.47	síntese	28	9.16
multivariada ✚	312	9.45	aprofundar	51	8.46	análise e síntese		
análise multivariada ,			aprofundar a análise			compreensão	32	9.00
exploratório ✚	319	9.42	possibilitar	74	8.39	análise e compreensão		
análise fatorial exploratória			facilitar	50	7.87	avaliação	35	8.73
detalhar ✚	268	9.16	facilitar a análise			a análise e avaliação		
uma análise detalhada			focar	31	7.87	apresentação	20	8.60
crítico ✚	341	9.12	desenvolver	74	7.72	análise e apresentação		
uma análise crítica			utilizar	77	7.67	observação	14	8.12
confirmatório ✚	227	9.01	dificultar	34	7.67	regressão	13	8.11
análise fatorial confirmatória			empreender	24	7.64	descrição	11	7.82

Figure 4-1 Partial word sketch of the noun *análise* ('analysis') in the *Corpus de Português Escrito em Periódicos* (CoPEP)

The Sketch Engine is a development of the word sketch software to encompass other languages – a frequent request of many a lexicographer working with languages other than English. Furthermore, it grew to be a corpus tool "both in the sense of

[135] I am thankful to the University of Ljubljana for receiving me for a Short-term Scientific Mission (COST Action IS1305; Grant number COST-STSM-IS1305- 210216-071459) under the supervision of Dr. Iztok Kosem and for giving me access to these two programs. I would like to thank Miloš Jakubíček for reducing the price of the academic licence so that I could continue my research after the Scientific Mission in Slovenia.

'corpus query tool' and in the sense of 'corpus web service'" (Kilgarriff et al., 2014, p. 34), which means that besides a comprehensive list of functionalities for corpus query, it provides preloaded corpora for a great number of languages as well as tools for corpus building and management. Currently, Sketch Engine offers more than 400 preloaded corpora in over 90 languages, with sizes reaching up to 20 billion tokens[136].

In addition to word sketch, some of the other functions that are available are KWIC Concordancer, Thesaurus, word sketch difference, Word lists, terms extraction, GDEX (Good Dictionary Examples), TickBox Lexicography, longest-commonest match (Kilgarriff, Baisa, Rychlý, & Jakubíček, 2015), parallel corpora building, and WebBootCat, to name but a few. I refer to Chapter 5 for a full description of the functionalities used in this thesis.

According to Kilgarriff (2006, p. 7), a dictionary writing system is "a piece of software for writing and producing a dictionary. It might include an editor, a database, a Web interface and various management tools […]". The iLex DWS, which is used in my project, combines a database with an editor.

iLex is the same tool used by the Slovene team. The decision to employing it in my project was based on three significant advantages. Firstly, an XML schema had already been customised and loaded, exempting me from having to hire a computational linguist to write one for my project. Secondly, I was granted access to this software and could download it to my personal computer, without any cost. Lastly, I did not have to go through experimenting different tools until finding the most qualified, which saved me a great deal of time.

4.2.1.2 Corpus

As previously mentioned, the target users of DOPU are university students who speak BP and EP and attend Portuguese-medium institutions, who thus need lexical support for both production and comprehension of academic texts. Given this profile, all the information that DOPU portrays must be obtained from texts that reflect the way language is used by expert writers from Brazil and Portugal producing texts in academic settings in different areas of knowledge. As shown in Chapter 3, at present there are no

[136] See the Sketch Engine web page for further information: https://www.sketchengine.co.uk/

corpora meeting that special demand. Thus, it was decided that a corpus had to be compiled from scratch.

Table 4.4 below shows the corpus compilation criteria laid down for meeting the lexical information envisaged for DOPU to portray:

Table 4.4 Criteria for corpus building

Lexical information in DOPU	Criteria for corpus compilation
Exemplary and typical academic Portuguese	Source: qualified academic texts
Portuguese as a pluricentric language	Balance: 50% of Portuguese texts, 50% of Brazilian texts
Representative of varied disciplines	Coverage: different areas of knowledge
Current language use	Period: synchronic
Comprehensive nomenclature Rich entries	Size: as large as possible

In addition to criteria determined by the aim of DOPU, there were also some working restrictions that had to be overcome, namely: a) lack of a team of professionals[137] – computational scientists, corpus linguists, lexicographers (Klosa, 2013); b) no budget allocation for personnel, computer technology (hardware and software) and copyright; c) limited time for data acquisition, copyright issues handling, and digitization – should this latter method be adopted.

A solution was found that seemed to cater for both corpus textual compilation criteria and corpus building working conditions: to download texts from free, peer-reviewed internet journals published in Brazil and Portugal. An extremely reliable source that attends to all these conditions is the SciELO (Scientific Electronic Library Online) platform[138], so it was decided that all texts in Portuguese from all journals in each national collection of the SciELO would be downloaded.

It is important to mention, however, that it was soon discovered that the Brazilian SciELO collection is much larger than the Portuguese collection. The main consequence of such a size difference was that, since one of the criteria for corpus building was equivalence between BP and EP, determination of the total size of the

[137] "Corpus linguists have to work on the corpus design together with lexicographers; computational linguists should be responsible for mapping the data structure upon consultation with lexicographers." (Klosa, 2013, p. 520). Although I did not have the support of a whole team of professionals, tasks that required advanced programming skills were performed by a colleague, José Pedro Ferreira.

[138] http://www.scielo.org

corpus was set by the smallest group of texts, i.e. that from Portugal. As McEnery et al. (2006, p. 71) have explained, "In building a balanced corpus according to fixed proportions (…) the lack of data for one text type may accordingly restrict the size of the samples of other text types taken".

The *Corpus de Português Escrito em Periódicos* (CoPEP) ('Corpus of Portuguese from Academic Journals') contains over 40 million words, is balanced between BP and EP, and covers six areas of knowledge. For this design, CoPEP was uploaded into the Sketch Engine, where it was POS-tagged with Freeling v.3 tagger. Comprehensive details on corpus compilation and post-processing are given in Chapter 6.

4.2.1.3 Sketch grammar

At the heart of the procedure of automatic extraction of data from the corpus are word sketches (see 4.2.1.1 above), which is a feature of the Sketch Engine tool (Kilgarriff et al., 2004) that provides a summary of the grammatical behaviour of a word. In order to build word sketches, two conditions have to be met. One is a POS tagged corpus (see 4.2.1.2 above), and the other is sketch grammar.

Sketch grammar is a file with grammatical relations, or gramrels, and processing directives for the Sketch Engine system to compute different types of relations through statistics calculations. The data obtained with these computations then form the basis of the word sketch feature in the Sketch Engine. Sketch grammars devised for POS-tagged corpora use regular expressions over POS-tags to find matches for grammatical relations. Queries are written in Corpus Query Language (CQL), with attribute-value names following the tagset used for corpus tagging.

Evaluation of the default sketch grammar provided by the Sketch Engine revealed several problems, indicating the need to develop a new sketch grammar for academic Portuguese. A decision was made to evaluate other existing sketch grammars for Portuguese, in order to determine whether they, or their parts, could be used for my purposes.

The overall conclusion was that neither of these sketch grammars could be used for data extraction, but rather that a completely new sketch grammar for academic Portuguese would need to be developed.

Devising the first sketch grammar was conducted on a provisional version of CoPEP and mostly focused on a test of writing the grammatical relation queries and an evaluation of the performance of these queries. Once the final version of CoPEP was ready, further improvements were implemented. A full account of a sketch grammar for academic Portuguese is given in Chapter 7.

4.2.1.4 Good Dictionary Examples configurations

As mentioned earlier, besides grammatical relations and collocates, this new procedure of data acquisition also automatically extracts examples from the corpus. For that, another feature of Sketch Engine is used, the GDEX function.

GDEX stands for Good Dictionary Examples (Kilgarriff et al., 2008) and is an automated method for finding examples in a corpus. The system is instructed to go through concordance lines and evaluate their components, assigning awards and penalizations according to predetermined parameters, producing a final score between 1 and 0. Those lines that score best (closer to 1) are placed first in the list of results, thus facilitating lexicographer's tiresome process of examples selection.

The first use of automatic extraction of good dictionary examples (GDEX) was in the context of the preparation of the *MacMillan English Dictionary for Advanced Learners* (2002, 2007) (see section 3.2.3). Since then, several other languages have developed GDEX configurations for automatic extraction of examples for lexicographic and language teaching purposes, including Slovene (Kosem, Husak, & McCarthy, 2011) and Estonian (Koppel, 2017). For Portuguese, a general configuration has been developed by the Sketch Engine team and is currently the default GDEX configuration available on the tool.

However, for automatic extraction of data performed in this PhD research, GDEX configurations should be fine-tuned to meet not only the characteristics of CoPEP, but also the purposes of these examples, given the target users of DOPU.

One of the challenges of working on a pioneering Portuguese lexicographic project such as DOPU is the lack of a benchmark against which parameters for resources creation can be measured and judged. It thus follows that, unlike the earlier presented GDEX configuration for English, Slovene, and Estonian, whose heuristics

were built on an analysis of existing manually validated good and bad examples, development of GDEX configuration for Portuguese involved finding alternatives to work around the absence of a representative model.

Chapter 8 shows the process of GDEX configurations development for academic Portuguese in detail.

4.2.1.5 Extraction

This phase reports on decisions taken regarding the preparation for the procedure of data extraction.

The point of departure followed the steps as described in Kosem et al. (2013) and Gantar et al. (2016) and a modified version of the Slovene API script. However, as preparation began, it became clear that further adjustments in the procedure would have to be made. Some characteristics that are exclusive to CoPEP, namely, texts from two national varieties of Portuguese and four different spelling norms,[139] resulting from the coexistence of three different spelling rules within the time span of CoPEP, posed unexpected challenges for data extraction.

On a positive note, analysis of two new features in the Sketch Engine suggested that lexicographers' work could be facilitated with their addition to the original procedure, leading to a decision to experiment this alternative.

The first addition was meant to attend to the demand of equal representation of the two varieties of Portuguese, especially as I was trying to assign variety labels not only to headwords, but also to collocations, and if relevant, to grammatical relations. I decided to extract data from both subcorpora varieties separately, and also add statistics on grammatical relations and collocations from the whole corpus.

While the first addition was especially developed for CoPEP, the second was actually language non-specific, i.e. it can be used for automatic extraction of other languages. It consisted of the inclusion of additional information provided by the

[139] In Chapter 1 the issue of the different orthographic agreements of the Portuguese language was succinctly touched upon. Texts produced before 2009 followed either FO43 or AO45 orthographic norms, the former governing Brazil and the latter, Portugal. Between 2009 and 2015, which was a transitional period for adaptation to the latest treaty, the AO90, which was ratified by the Member States of CPLP (except for Angola and Mozambique), both norms were accepted in each country. As the CoPEP period of time is from 2000 to 2016, that means there are texts reflecting the application of four spelling rules for text writing. A detailed account is given in Chapter 6.

clustering and longest-commonest match functions in the Sketch Engine. This information was added to the data after the extraction, at a post-processing stage. The main aim was to assist lexicographers in grouping collocates, and in identifying multi-word expressions, as well as facilitating the detection of incorrect information.

Firstly, an experiment was carried out to evaluate the method of automatic extraction of data from CoPEP as a means to provide lexical content that would serve as a basis for compiling entries of DOPU. At that time, data information from a sample of lemmas was extracted from a provisional version of CoPEP, employing preliminary versions of Sketch Grammar and GDEX configurations.

Evaluation of the output was positive, revealing that the procedure with the two additions mentioned above could be successfully employed for data acquisition for compiling entries of DOPU. It was also concluded that improvements in CoPEP, Sketch Grammar, and GDEX configurations should enhance the quality of the data extracted. I refer to Chapter 6 for CoPEP post-processing, Chapter 7 for Sketch Grammar devising, and Chapter 8 for the tweaking of GDEX configurations.

After implementation of these improvements, a second extraction was performed. Chapter 9 gives a comprehensive description of the preparation for the procedure, including outcomes of each step, as well as report on both the experiment and the second extractions.

4.3 Candidate headword list building

Having decided on the method for knowledge acquisition from the corpus in order to serve as a basis for entry writing, the next stage referred to DOPU macrostructure. As Hartmann and James (1998, p. 91) have explained, the central component of the macrostructure of a dictionary is a word list, accompanied by the optional front, medium and back matters. In this thesis, focus was given to the word list – also called headword list, nomenclature, or A-Z list – while the case for additional lexicographic-related informative material was only briefly touched upon, as will be shown in Chapter 10. Thus, the objective of this phase was to plan how to build the candidate headword list, pointing out the decisions that should be made.

For organizational purposes, I will begin by touching upon issues related to the delimitation of vocabulary in the headword list, then move on to specific matters at the level of lexical entries.

Before moving on, however, some terminological clarification is required regarding the definition of some key concepts[140] for the lexicographic work developed in this thesis, which are summarised in Table 4.5.

Table 4.5 Key concepts

lemma	the stemmed form of the word e.g. "cat" is the lemma for the word form "cats"
lexical item	the abstract unit of the lexicon (Lipka, 1990, p.73)
lexical unit	the union of a lexical form and a single sense (Cruse, 1986, p. 77)
token	any word or punctuation
word	used in a general sense when there is no need for the specific technical terms and no ambiguity arises (Lipka, 1990, p.73)
word-form	the orthographic form of the word. It encompasses the inflected forms of verbs, adjectives and nouns, and the derivative forms of adjectives, adverbs and nouns.

4.3.1 Delimitation of vocabulary

First and foremost, as a corpus-driven dictionary, DOPU headwords would have to be extracted from CoPEP. Although the advantages of employing corpora in dictionary-making have already been presented in Chapter 3, it is worth recalling their specific contribution to dictionary macrostructures.

One important role that corpora play in the process of headword list selection can be seen as that of a content gatekeeper, assuring that the lexical items granted headword status are consistent with the stretch of language that the dictionary aims to cover. This is why CoPEP was especially compiled for this project. By providing evidence of the language that DOPU target users tend to encounter in the context of routine activities of reading and writing in higher education, it is expected that relevant

[140] As is widely known, there is a large body of studies approaching the lexicon from various theoretical perspectives in different areas of language studies (e.g., Computational Linguistics (Kilgarriff, 1997); Corpus Linguistics (Sardinha, 2004; Sinclair, 1991); Discourse Analysis (Hoey, 2005); Lexical Semantics (Cruse, 1986); Lexicography (Atkins & Rundell, 2008; Biderman,1984a; Svensén, 2009; Zgusta,1971); Lexicology (Correia & Lemos, 2009; Lipka,1990; Vilela, 2002); Linguistics (Lyons, 1981), to name but a few). Given the applied nature of this thesis and, consequently, the need for the development of a series of new resources and tools, crucial concepts were chosen according to their operational character, without restriction to one determined theory. This is not surprising, as lexicography is referred to as a practice that is "far too varied and multi-faceted to be covered by a single theory" (Bogaards, 2010, p. 318).

vocabulary is covered. It can be said that employing CoPEP for the compilation of the DOPU headword list puts into practice the recommendation by Gouws' (2003, p. 38):

> The macrostructure of any general dictionary (both monolingual and bilingual) or dictionary dealing with a language for special purposes (LSP) has to reflect that section of the lexicon of the language relevant to the scope of the dictionary. This implies that all the types of lexical items prevailing in that section of the lexicon have to be included in the macrostructure.

Moreover, since the Cobuild project (Sinclair, 1991; Moon, 2007; see Chapter 3 for a detailed account), it has been known that corpora are at their best when revealing what is common and what is not in language use. Indeed, de Schryver (2013, p. 1385) points out that corpora "make it possible to separate the frequent and average from the one-offs; to distinguish the typical from the oddities". Considering that DOPU is planned to describe typical uses of academic Portuguese, frequency was decided to be the main criterion for inclusion or omission of lemmas in a headword list.

Experiments with CoPEP for defining a suitable minimum frequency cut-off point should be performed. That means that a frequency-ordered lemma list extracted from corpora should be the point of departure for candidate headword building (see Biderman, 1984b, 1998; Gouws, 2003; Kosem, 2010; Welker, 2004).

At this point, attention should be drawn to a beneficial implication of the characteristics of DOPU (non-commercial product and online medium) to macrostructural decisions: the size of the headword list need not comply with some externally-imposed limitations. For instance, on the one hand, as a dictionary developed within an academic context, DOPU is exempted from having to follow market-oriented principles typical of commercial dictionaries, including definitions of headword list content and size. On the other hand, as an online dictionary, the number of headwords does not have to be restricted to a closed set pre-defined by space constraints, which tends to be the norm with print dictionaries.

Evidently, having some leeway to determine the nomenclature size does not mean that anything goes. This is why laying down rigorous selection criteria is paramount to ensure the quality of the candidate headword list. Thus, besides corpus frequency, other decisions that need deliberation involve treatment of headwords, i.e. a

homonymic or polysemic solution, and the inclusion or omission of words in reference to the corpus.

In general terms, homonymy concerns two different lexical items sharing the same lexical form, while polysemy refers to one lexical item covering more than one lexical unit. The importance -and most of all, the challenge- of distinguishing homonymy from polysemy have been referred to in a number of studies (e.g. Atkins & Rundell, 2008; Biderman, 1984b, 2005; Cowie, 2001; Cruse, 1986; Landau, 2001; Lyons, 1981; Svensén, 2009; Zgusta, 1971; to name but a few). In the context of dictionary-making, determination of homonymy and polysemy is particularly significant when lexicographers have opted for separate treatments of each case. In such a scenario, homonyms are included in the dictionary as different headwords, thus interfering with the macrostructure. Polysemic lexical items, on the other hand, are accorded one headword and the different senses (i.e. the different lexical units) are dealt with within that entry, making this phenomenon a microstructural issue.

A typical example[141] of homonymy in English is the case of *bank*, with *bank*^{1} 'financial institution' and *bank*^{2} 'river bank'. In Portuguese, the equivalent of 'bank' is also a homonym: *banco*^{1} 'financial institution', whereas *banco*^{2} 'seat for more than one person'.

Polysemy is characterised by semantic relatedness. So, taking the lexical item *bank*^{1} above, it contains the lexical unit *bank 1*. 'Financial institution' and *bank 2*. 'Storage place (blood bank)'. The same applies to *banco*^{1} in Portuguese (e.g. *Banco do Brasil* and *banco de sangue*).

As to inclusion and omission of words in reference to the corpus, it is worth highlighting a very important question in the case of Portuguese, which has an official regulation of the orthography of words: should the candidate headword list include words that occur in the corpus but are not attested in VOC (see Chapter 1)? This is a very complex issue that requires further examination of topics such as language norm and language use, the role of dictionaries in general and of a dictionary of academic Portuguese in particular, and official regulatory norms, to name but a few.

[141] Definitions of the examples of homonymy and polysemy (in the following paragraph) are not provided, but rather a simple reference to the lexical field.

4.3.2 Lexical entries

With regard to lexical entries more specifically, it was clear that proper deliberation on the types of lemmas that should be given headword status was crucial for building a theoretically-sound, coherent nomenclature of DOPU. These decisions, as is known, depend on the understanding of what a word is. Nonetheless, the answer is far from simple.

A suitable theoretical perspective for identification, classification, and description of Portuguese words concerning *form* is the framework delineated by Correia and Lemos (2009) to address processes of word formation in the Portuguese system. The authors have referred to the concept of graphic word, i.e. the prototypical written word understood as a string of characters delimited by two blank spaces, to propose two other types of words: those with a larger dimension and those with a smaller dimension than the graphic word.

Some cases falling into the first group are:

- (Noun) compounds (*sala de jantar*, 'dining room'; *casa de banho*, 'bathroom')
- Prepositional (*por cima de*, 'over'), conjunctional (*visto que*, 'given that') and adverbial (*de cor*, 'by heart') compounds.[142]

As to words smaller than the graphic word, Correia and Lemos (2009, p. 7) explained that they refer to non-autonomous words. These are words with stable meanings, belonging to morpho-syntactic categories, but that can only occur as elements in the formation of other words, and never as a free form, such as *psic-*.

Finally, the authors addressed idioms, explaining that a lack of compositional meaning suggests that they are in speakers' lexical memory together with other words, thus deserving proper description in dictionaries.

[142] Correia and Lemos used the term '*locução*' to refer to a lexical item composed of more than one word that functions as one specific word category. As shown above, the example given for *locução preposicional - por cima de* - is composed of preposition+noun+preposition, functioning as a preposition in a sentence. *Compostos* (compounds) are used for nouns and, rarely, adjectives, and invoke the process of word formation underlying these lexical items, namely, lexicalization of phrases. For an approximate conceptual correspondence in English, I adopted the terminology in Atkins and Rundell (2008, pp. 169-170), who employ the term 'compound' for both '*compostos*' and '*locução*' – they give 'in spite of' as an example of a compound preposition. However, the authors emphasise that the interest of lexicographers is in nouns, adjectives and verb compounds (mostly, phrasal verbs).

From this perspective of form, finding single words in a corpus seems to be a fairly clear-cut process: the blank spaces determine the unit. That means that meaning analysis is a secondary step. This is why computers are used for the automation of word identification.

By contrast, "identifying uniquely meaningful multiword expressions is a task that poses particular problems"(Hanks, 2010, p. 587). This is because "precise criteria for drawing the line between free and fixed components are simply impossible to set", according to Atkins and Rundell (2008, p.167).

Correia and Lemos (2009) also acknowledged that identification of phrasal compounds is extremely difficult. The authors presented phrasal compounding[143] as one of the possible processes of word formation in the Portuguese language. That is, phrasal compounds are the result of a process of lexicalization of phrases. In this context, Correia and Lemos (2009, p.40) defined lexicalization as

> The process by which certain units built in other components of the grammar (syntactic, morphological, discursive) are turned into lexical units that become part of the language, functioning as fully fledged lexical units on their own right[144]. (my translation).

Atkins and Rundell (2008, pp. 164, 166-179) included compounds as one type of multiword expression, which is the term that they employ that "covers all the different types of phrases that have some degree of idiomatic meaning or behaviour" (p. 166). Under this class of lexical items, the authors also include:

1. Fixed and semi-fixed phrases (transparent collocations, fixed phrases, similes, catch phrases, proverbs, quotations, greetings, phatic phrases)
2. Other phrasal idioms
3. Phrasal verbs
4. Support verb constructions

[143] *Composição sintagmática* in Portuguese.

[144] "*o processo pelo qual determinadas unidades construídas em outras componentes da gramática (sintáctica, morfológica, discursiva) se transformam em unidades lexicais […] que se fixam na língua, passando a funcionar como unidades lexicais de pleno direito*".

Under “other phrasal idioms”, there are some cases which are very difficult to distinguish, not fitting any of the other categories above. Atkins and Rundell (2008) have given some suggestions on how to analyse them: lexicographers should verify if an MWE has one or more of the following properties (the authors draw attention, however, to the fact that “no idiom has them all” (p.168)):

- The meaning is more than the sum of the parts
- The wording is never entirely fixed
- There are syntactic restrictions upon the idiom’s behaviour, in that it undergoes only limited grammatical transformations
- The idiom shows morpho-syntactic flexibility, allowing inflections, agreement of possessives, and so on.

These tests are valid for Portuguese as well, as is shown in Biderman (2005).

It should not be forgotten, however, that the actual lexicographic work of multiword expressions analysis for the purpose of selecting candidate headword lists will not be done in this thesis. The nature of this task demands painstaking analysis of the lexical item in its textual context, which is part of the lexicographic routine performed during database entry compilation.

Finally, in addition to identification of single lexical items and MWEs, decisions were made as to which types of words should be granted the status of candidate headword, for instance, acronyms, proper nouns, derived forms, loan words, etc.

Taken together, these are the main factors that need to be accounted for when building the candidate headword list of DOPU. Decisions will be made by applying some of the theoretical contributions presented here to the analysis of CoPEP, as can be seen in Chapter 10.

4.4 Entry compilation

With a clear target user in mind, a defined method of extraction of lexical knowledge, and an outline of the candidate headword list, it was time to go over the factors involved in decision-making about “the central part of the lexicographer’s work,

i.e. the construction of the entry with the definitions" (Zgusta, 1971, p. 355). These were:

- Microstructure (see 4.4.1 below)
- Data analysis (see 4.4.2 below)
- Secondary sources (see 4.4.3 below).

4.4.1 Microstructure

The microstructure of a dictionary concerns the elements carrying information about the headword in the entry and how they are organised. Thus, the initial step in my research was to make a record of different types of information. For that, I resorted to renowned dictionary-makers (e.g. Atkins, 2008[1992/3]; Atkins & Rundell, 2008; Kiefer & van Sterkensburg, 2003; Landau, 1981; Svensén, 2009; Zgusta, 1971), making a record of the following categories:

- Variant form
- Pronunciation
- Syllabification
- Word class
- Inflection
- Etymology
- Syntactic structure
- Semantic pattern
- Usage
- Synonym/antonym
- Collocation
- Multiword expression
- Definition
- Examples
- Sense menu

In computer-assisted lexicographical projects, it is customary to draft a number of entries during the planning stage (Atkinson & Grundy, 2006). It has been shown that data analysis indicates what additional information should be included in the entry to

better cater for the whole description of the headword (Grundy & Rawlinson, 2015, p. 567). This leads to the determination of an entry layout, which in turn informs the design of the dictionary type documentation (DTD). Grundy and Rawlinson (2015, p. 567) have provided a succinct and instructive definition of DTD:

> The DTD lays down the underlying structure and associated rules for the building of the dictionary text. (…) The DTD establishes the hierarchical relationship between the various elements of the entry. It is thus the DTD that provides the underlying XML structure of the dictionary…

In the case of my project, where automatic extraction of knowledge from the corpus and import into Dictionary Writing System (DWS) is the method adopted for dictionary-making, I used iLex (as mentioned in 4.2.1.1 above) and a pre-defined project schema.[145] That means that the type of microstructural decisions made in my thesis did not lead to the devising of extra elements in the XML Schema, but rather in defining which ones should be included in a prototypical entry of DOPU.

Thus, in Chapter 11 I delve into each one of these types of information, making decisions about how they should be handled in DOPU and occasionally illustrating these decisions with provisional entries.

4.4.2 Lexical analysis

The key innovative aspect of the method of automatic extraction from the corpus and import into DWS is the fundamental change in the lexicographer's core work, namely, meaning analysis, which now begins directly in the DWS.

In the context of DOPU design, entries were populated with automatically extracted information on headword, word class, frequency, grammatical relations, clustered collocates and examples for each collocate. The process of lexical analysis for entry writing thus involved the following steps.

Firstly, collocations under all grammatical relations of the entry were read, leading to initial sense differentiation. In Aktins and Rundell's (2008) terms, this is part

[145] As informed in the iLex design manual, "iLEX use XML Schema as defined by W3C. The schema used for a project is specified in the iLEX project definition by XML Schema. XML Schema is written as XML documents (as opposed to DTDs). XML Schema can therefore be created and edited within iLEX itself".

of the process of sense disambiguation: grouping collocates under different lexical units.

Next, corpus lines that had been automatically selected from CoPEP as good examples candidates were analysed in order to confirm if they matched the associated lexical unit. Bad examples and wrong collocates were deleted. Senses elements were added to the entry and indicators were filled with a succinct definition. Collocates and examples were sorted under the lexical units.

Next, for each lexical unit, a fine-grained lexical analysis was performed to obtain information about collocational and colligational behaviours, in addition to a description of semantic patterns. Analysis of metadata available through entry XML visualization provided extra relevant information, like the area of knowledge of the examples and frequency in the corpus.

Database entry compilation also involved additional actions, such as determining sense ordering, the writing of lexical units' definitions, use of some functions of the Sketch Engine for complementary information, and comparison with secondary sources (see next section), to name but a few.

At this point, it should be stressed that the whole process of entry compilation just described refers to elements in the entry database. As is known, the common practice in lexicographical projects is to include much more information than what will be displayed for the final user (Atkins & Rundell, 2008; Bergenholtz & Nielsen, 2013).

In this thesis, a decision had to be made in which the application of the method of automatic extraction of data from the corpus and import into DWS was favoured over the compilation of dictionary-entry samples. As has already been explained, this required development of specially-built resources and tools for Portuguese.

4.4.3 Sources

The primary source for entry compilation was CoPEP. However, at times it was necessary to resort to secondary sources. This is part of the lexicographer's work, as Čermák (2003, p. 19) has stated:

> lexicographers always consult other dictionaries or previous editions of the same dictionary. With their main goal being

> verification of their own definitions and the general treatment of an entry, they specifically look for omissions, changes and new features or words not recorded before or recorded elsewhere. When in need of more information and data support, they may specifically consult their corpus, if any, use specialised dictionaries, indexes or encyclopaedias (in the case of terms, usually) or resort to other techniques.

Secondary sources used for reference were the following monolingual dictionaries of Portuguese, namely:

- *Dicionário Aurélio da Língua Portuguesa*, Brazilian Portuguese (print and CD, 2010). Henceforth *Aurélio*.
- *Dicionário de usos do Português do Brasil* (print, 2001), by Francisco S. Borba. Henceforth *Borba*.
- *Dicionário eletrônico Houaiss*, Brazilian Portuguese (CD, 2009). Henceforth *Houaiss*.
- *Dicionário da Língua Portuguesa Contemporânea* (2001). Henceforth *Academia*.
- *Grande Dicionário da Língua Portuguesa*, European Portuguese (Porto Publisher, 2016). Henceforth, *Porto*.
- https://www.infopedia.pt/ , which is the free online dictionary of Porto Editora. Henceforth, *Infopedia*.
- https://www.priberam.pt/dlpo/ . Henceforth, *Priberam*.

Additionally, the online dictionaries available in the *Portal da Língua Portuguesa* were also consulted, i.e. dictionary of gentiles and toponymy, dictionary of loan words, and dictionary of deverbal nouns.

Another important lexical reference used was VOC (*Vocabulário Ortográfico Comum*; 'Common Orthographic Vocabulary of the Portuguese Language'),[146] with more than 300,000 words from different Portuguese varieties. This vocabulary consists of the official spelling for the variety used in each CPLP country, meaning that it represents the orthography according to the recently implemented orthography reform (see Chapter 1).

[146] http://voc.cplp.org/index.php

For grammatical issues related to lemmas, the following grammars were consulted:

- *Gramática do Português*, vol. I and II. (Gulbenkian Publisher, 2013)
- *Moderna Gramática Portuguesa* by Evanildo Bechara
- *Modern Portuguese* by Mário Perini
- *Nova Gramática do Português Brasileiro* by Ataliba Castilho

4.5 End-user interface

Online dictionaries should take advantage of the condition of being inherently connected to the World Wide Web, as well as from the new possibilities that the digital medium allows.

One of the assets that has been increasingly used is the customization of searches according to the user profile,[147] following, in a way, Sue Atkins' foresight, of which Varantola has reminded us:

> In her paper, Sue Atkins anticipates that, in the dictionary of the future, the function of customizing the dictionary will 'come into its own' (1996: 531). Future dictionary users would then be able to tailor the dictionary according to their individual user profiles. (Varantola, 2002, p. 31)

According to Varantola, this customization allows dictionary makers to focus on users' real needs. This is my aim with DOPU in the wake of other lexicographic projects whose users are also university students, like Granger and Paquot (2010a, 2010b) and Kosem (2010).

Granger and Paquot (2010a, 2010b) (see also Paquot, 2012) designed a dictionary-cum-writing aid – the *Louvain EAP Dictionary*[148] (LEAD) – for university students who are also English learners, with customizable search option, granting users with specially-composed entries to meet their needs in terms of mother tongue background and discipline of interest. A similar proposal is present in Kosem's (2010)

[147] For an analysis of four different types of current customization in Internet Dictionaries, see Trap-Jensen (2010).

[148] This dictionary is only available for members and students of the *Université catholique de Louvain*. https://www.uclouvain.be/en-322619.html

model of a dictionary of academic English, whose lexicographic project has never been taken on. Kosem's model is aimed at university students, both speakers of English as a mother tongue and as an additional language. Thus, the content of the entry is adapted to the user profile, according to three external characteristics of students (Kosem, 2010, p. 301): place of study, native language, and subject of study.

In a similar vein, search customization in DOPU should be based on roughly the same categories: language variety and area of study.

For language variety, there should be two methods of customization: automatic and manual. The former refers to identification of the country where the dictionary is being consulted, via the IP address, and automatic display of the interface (and the underlying database) only for that particular variety. Evidently, in cases where the dictionary is accessed from other countries than Brazil and Portugal, or if users want a different variety than the one coinciding with country of access, manual selection could be performed. Additionally, a default interface allowing for searches in both language varieties together might be set; should users be interested in only one of them, icons for each variety are clickable, restricting database searches and data display.

With regard to the subject of study, this is planned to refer mostly to sense ordering and examples presentation. The idea is to present a full entry for the search lemma, whose senses are ordered by frequency by default. However, in the menu, icons for areas of knowledge covered in CoPEP should be placed next to a sense that is domain-specific. By clicking on the icon, the user is taken directly to the definition of that sense, including examples.

Another advantage that digital support brings to DOPU end-user interface has to do with hyperlinks to images/gifs/videos, which are available on the internet and can help greatly to clarify entries' senses (see 4.1.2.2 above). Certainly, this strategy shall not be applied to all headwords given their different statuses. Thus, concrete nouns, action verbs, and proper nouns (especially technical words) are apparently good candidates for using hyperlinks.

4.6 Concluding remarks

After the presentation of the process of planning DOPU, it is now clear who the target users are, what the purpose of DOPU is, and how to go about developing it. All fundamental factors concerning the different stages of the lexicographical process for making DOPU have been discussed. Moreover, goals for each phase have been set and indications on how to reach them have been presented. With an envisaged outcome of DOPU in mind, I can now move on to the practical part of this project.

Part II

SET-UP FOR SEMI-AUTOMATED LEXICOGRAPHY

Introduction to Part II

As previously stated, a major part of the design of DOPU that I propose in this thesis is the adoption of the semi-automated approach to dictionary compilation, as originally proposed by Rundell and Kilgarriff (2011) and first implemented into lexicographic practice by Gantar et al. (2016) (see also, Logar, Gantar & Kosem, 2014; Logar & Kosem, 2013). In this approach, lexical data (grammatical relations, collocations, examples) are automatically extracted from the corpus according to predetermined criteria, and transferred to the dictionary writing system, where lexicographers then analyse, validate and edit the data to shape them into the final entry. As Gantar et al. (2016, pp. 218-219) have reported, the comparison of the manual approach (analysing and selecting relevant data in the corpus tool) with the semi-automated approach showed that the latter is more effective and time efficient, and streamlines the lexicographical process without reducing the quality of the information provided in the dictionary.

Part II of this thesis will account for the assembling of the requirements for carrying out the extraction and import into DWS, namely, software, a corpus, a sketch grammar, GDEX configurations, and parameter settings for extraction.

In Chapter 5, a comprehensive presentation of the Sketch Engine, the corpus tool which is at the heart of this process, is given together with an overview of the main function of iLex, the DWS that was adopted in my research. In Chapter 6, I will report on the compilation of a corpus especially put together to attend to the demands of DOPU. Chapter 7 presents the devising of a new sketch grammar for CoPEP. Part II ends with the description of the development of GDEX configurations especially tweaked taking into account CoPEP's characteristics.

Chapter 5 Software

In this chapter, I will introduce some of the functionalities of the Sketch Engine (Kilgarriff et al., 2004) corpus tool and the iLex (Erlandsen, 2010) dictionary writing system. These two programs were used for the procedure of automatic extraction of data from the corpus and import into DWS as originally created by the team of lexicographers developing the Slovene Lexical Database (cf. Gantar et al., 2016) and that was followed closely in this PhD research for designing DOPU. The reason for adopting the same programs in my project was to facilitate the adaptation of the extraction procedure to the particular characteristics of the DOPU lexicographic project (see Chapter 9 for a full account of the procedure).

5.1 The Sketch Engine

The Sketch Engine is undoubtedly the most sophisticated corpus tool currently available. A general overview was given in Chapter 3, so I now turn to the presentation of specific functions that are used for interrogating a chosen corpus. Due to space constraints, I will focus on the functions I used the most in my research, i.e. bearing in mind that my purpose with corpus analysis is to make a monolingual dictionary. This is very important as there are many other features for lexicographers working on bilingual dictionaries, for translators, for language teachers, and for terminologists that will not be introduced here. For more information on the Sketch Engine, I recommend Atkins and Rundell (2008, pp. 103-113), Thomas (2015) and the user guide in the Sketch Engine website,[149] which contains a very rich material attending to different users' needs.

The corpus chosen for this demonstration is the 40-million-word CoPEP, which was briefly touched upon in Chapter 4 and will be described in detail in Chapter 6.

After selecting a corpus – CoPEP – we are taken to a screen that looks like this (Figure 5-1):

[149] https://www.sketchengine.co.uk/user-guide/

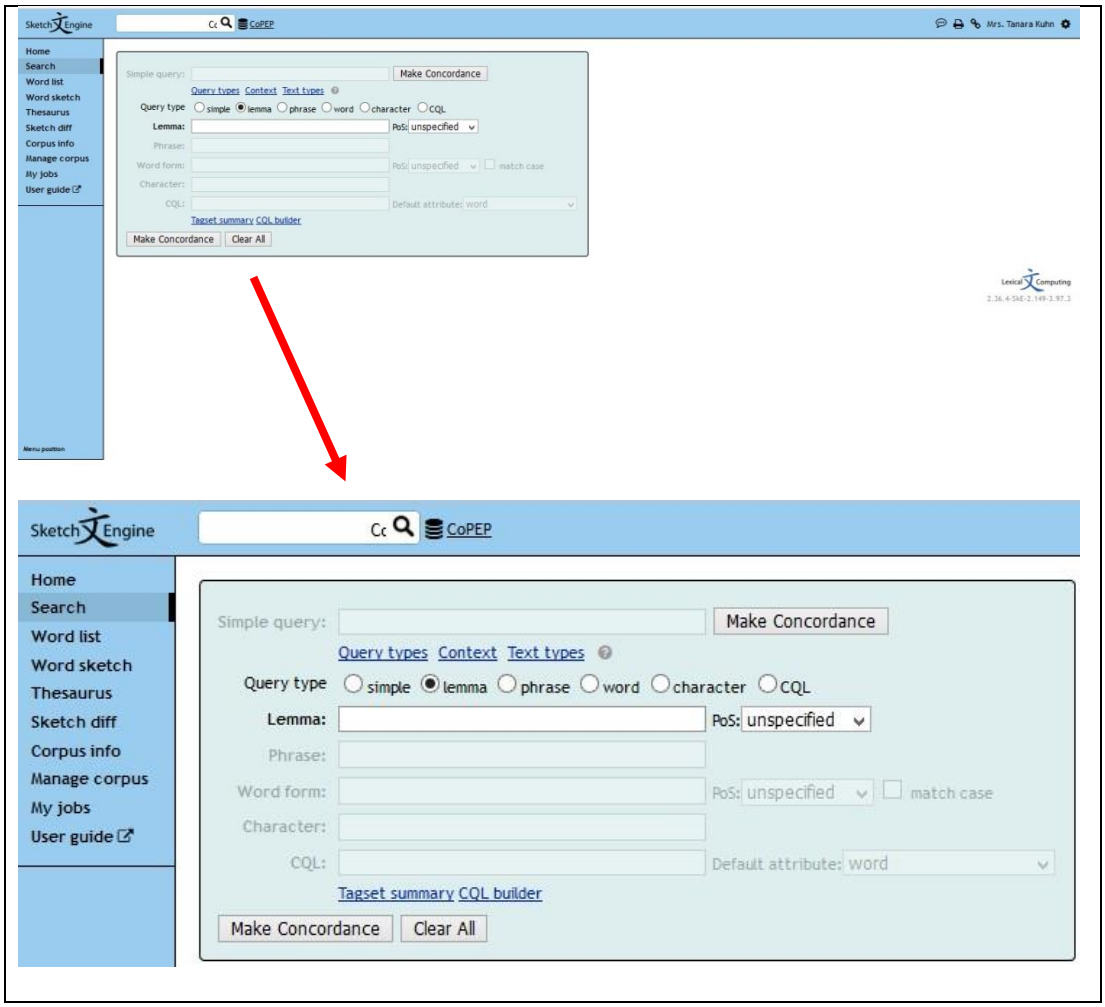


Figure 5-1 Initial screen in Sketch Engine

At the top of the screen, the name of the corpus in use is displayed next to a search box that allows for quick, simple lemma searches, which display KWIC (Keyword-in-Context) concordance results. On the left-hand side of the screen, the Sketch Engine functions displayed are *Search*, *Word list*, *Word sketch*, *Thesaurus*, and *Sketch Diff*, which will be described further in the following subsections. The remaining options refer to corpus tool navigation (*Home*, *User Guide*), display of corpus information (*Corpus info*), corpus management (*Manage corpus*) and verification of completion statuses of processing background jobs (My jobs), like the creation of a subcorpus (Figure 5-2).

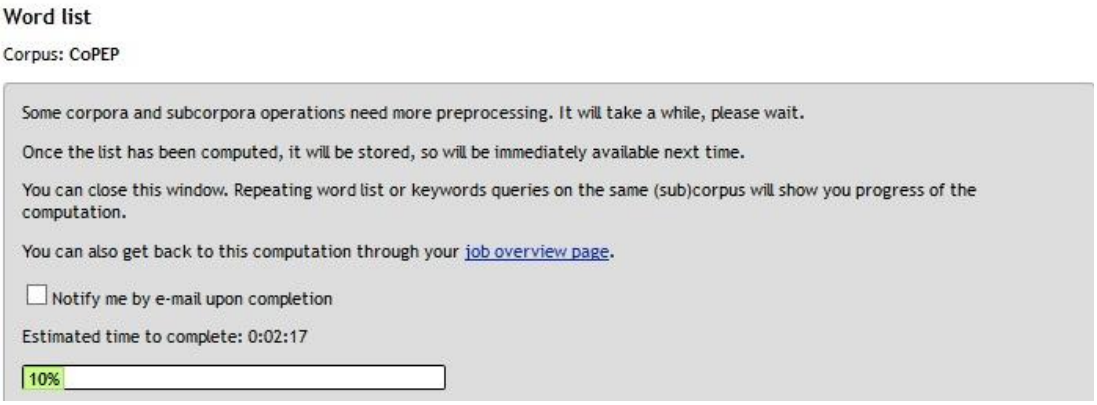


Figure 5-2 Background job processing

5.1.1 Search

Different kinds of searches can be carried out in the Sketch Engine, as can be seen in Figure 5-1 above. *Simple query* generates concordances of the keyword according to what is typed (lemma, word form, phrase). When clicking on *Query types*, other search options are offered: *Lemma* (which allows for POS specification), *Phrase* (more than one word, including a whole sentence), *Word form* (which allows for POS specification and selection of case matching), *Character*, and *Corpus Query Language* (CQL) with attribute choice, according to the corpus annotation (e.g., word, lemma, tag). More on using CQL for advanced searches will be described in Chapter 7. *Tagset summary* shows in a new tab the tagset used for tagging the corpus. In addition, *CQL builder* is a recent innovation that allows less experienced CQL users to create more advanced searches (Figure 5-3).

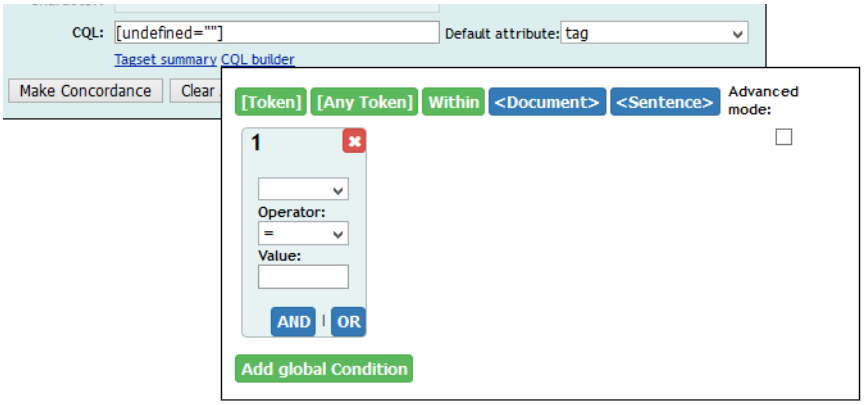


Figure 5-3 CQL builder

Context, which takes us to the second advanced search option, provides extra query filtering with regards to lemma and POS (Figure 5-4). For instance, I want to see which adjectives follow *introdução* ('introduction'). The searched word is typed in the *Lemma* field in the upper half of the pane; then, under *Context*, in the lower half of the pane, I define the search for only adjectives immediately following *introdução*. For this purpose, the POS filter is set to look for adjectives in a window of one token to the right of the keyword, or node, *introdução*.

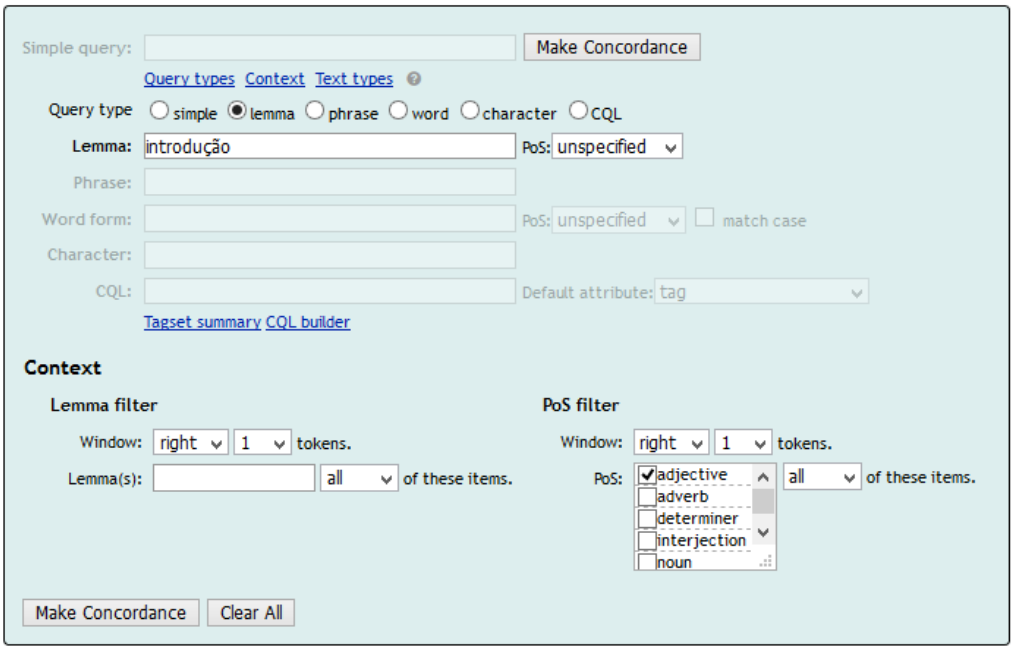


Figure 5-4 Advanced search with context filter

Figure 5-5 Search results screen

On the result screen (Figure 5-5), at the top of the screen one can see the keyword in red (*introdução*), the number of total occurrences of the keyword in the corpus (9,104 occurrences), the filter applied (adjective), the order of presentation (according to GDEX rank, which will be further explained below) and the hits per million of the construction *introdução*+adjective (2.7 million).

Finally, *Text types* searches depend on metadata annotation of the corpus. In the case of CoPEP, a series of relevant data were annotated to allow for advanced searches, for instance, language variety or great area of knowledge. I will not go into details here as a whole explanation will be given in Chapter 6. Figure 5-6 below shows a part of the kind of text type search available in CoPEP.

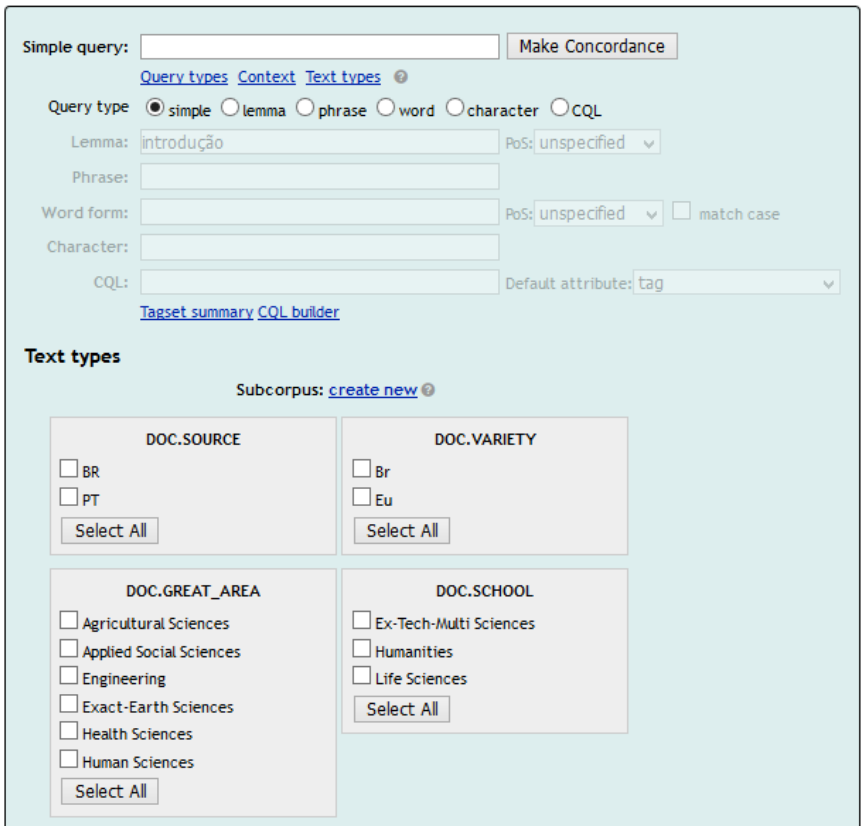


Figure 5-6 Text type advanced search

For the function *Search*, the default result screen displays KWIC concordances, with the search word centred and with some context to the left and to the right of the node.

In the result screen of a *Lemma* query for *metodologia* (Figure 5-6), one can see some extra information, which I had specially selected to be displayed by clicking on *View options* in the left menu.

In the centre of the concordance lines, next to the keyword, the POS-tag and the lemma are shown in grey. In blue, to the left of each line, the area of knowledge of the text with that occurrence is shown. Additionally, to the right of each line, there is an icon for copying the sentence to be pasted anywhere.

Figure 5-7 Result screen for a lemma search (*metodologia*)

From these KWIC concordance results, more sophisticated searches can be made.

5.1.1.1 Refining search results

One interesting query refers to the area of knowledge in which *metodologia* is used.

I first type *metodologia* in the *Lemma* search field, choose *noun* from the drop-down box POS attribute and hit *Make concordance*. This is the result screen (Figure 5-8):

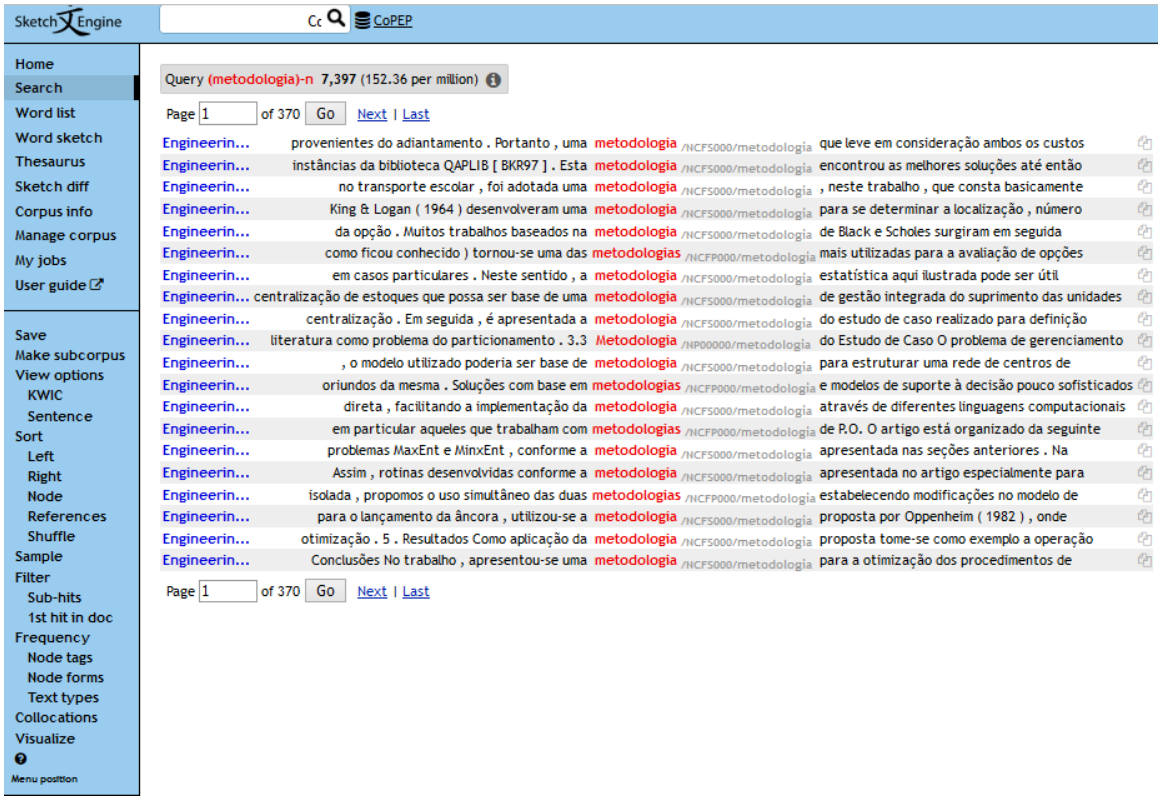


Figure 5-8 Result screen for lemma search (*metodologia*) with left menu

In the lower half of the left menu, there are many options to further work with the results. I will click on *Text types* under *Frequency*, in order to see the distribution of *metodologia* across the six areas of knowledge which are covered in CoPEP (Figure 5-9).

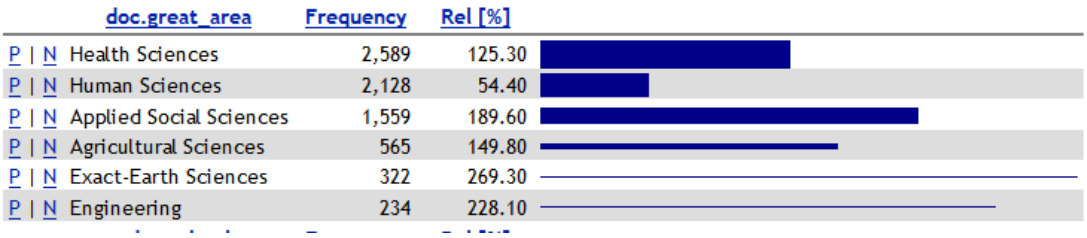


Figure 5-9 Distribution of the occurrence of *metodologia* according to areas of knowledge

As is known, areas of knowledge have different sizes in CoPEP; this is why raw frequency does not represent an accurate proportion. Instead, we should consider the relative text type frequency (Rel). This calculation compares the frequency of a keyword in a text to its incidence in the whole corpus. According to the documentation

in the Sketch Engine web page,[150] "The number is the relative frequency of the query result divided by the relative size of the particular text type".

Thus, the result above indicates that *metodologia* occurs almost 2.7 times more often in texts from the area of Exact-Earth Sciences than in the whole corpus, while the keyword is used only 0.5 times more often in the Human Sciences than in the entire corpus.

Another possible query concerns the inflection of number of *metodologia*: is there any significant difference in frequency between singular or plural forms? Going back to the KWIC concordances results screen (Figure 5-8), I will now click on *Node forms*. This is the result:

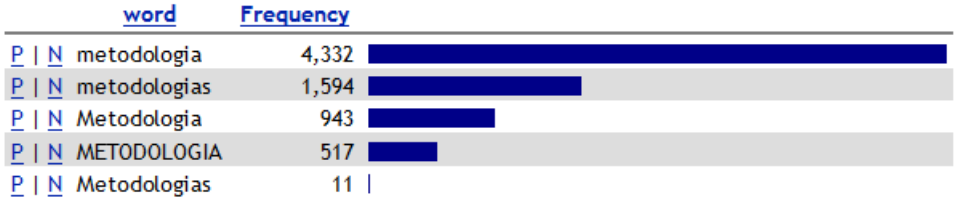


Figure 5-10 Distribution of node form of *metodologia*

The result shows more occurrences of singular than plural forms of *metodologia*. However, further investigation of lexical behaviour considering contexts of use of each form is required to determine in which senses the difference of inflection of number should receive a usage note. For instance, those who are familiar with academic papers know that *Metodologia* - with a capital letter and in singular form - might indicate the title of a paper section. Further analysis of neighbouring words and larger stretches of text should provide an answer to that hypothesis.

From the same KWIC concordance results for *metodologia* (Figure 5-8), it is possible to quickly verify whether pre-nominal attributive adjectives are used with the keyword. For that, I will click on *Frequency* in the lower half of the left menu in order to use the Multilevel frequency distribution function Figure 5-11:

[150] https://www.sketchengine.co.uk/my_keywords/relative-text-type-frequency/

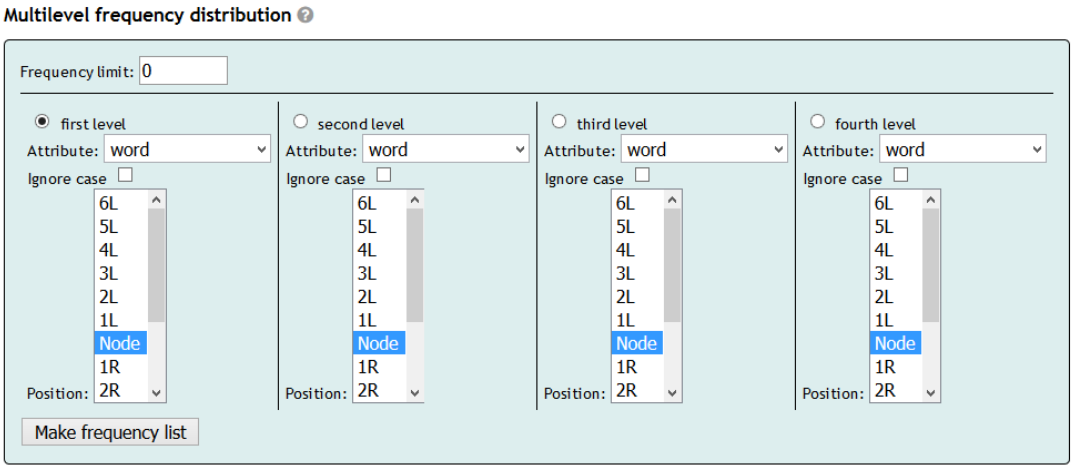


Figure 5-11 Multilevel frequency distribution concordance sorting (*metodologia*)

Next, I want to know what the POS of the words preceding the node are. Under *first level*, I choose 1L (one token to the left of the node) and select *tag* in the drop-down box for *Attribute*. Next, under *second level*, I choose Node and select *lemma* (Figure 5-12).

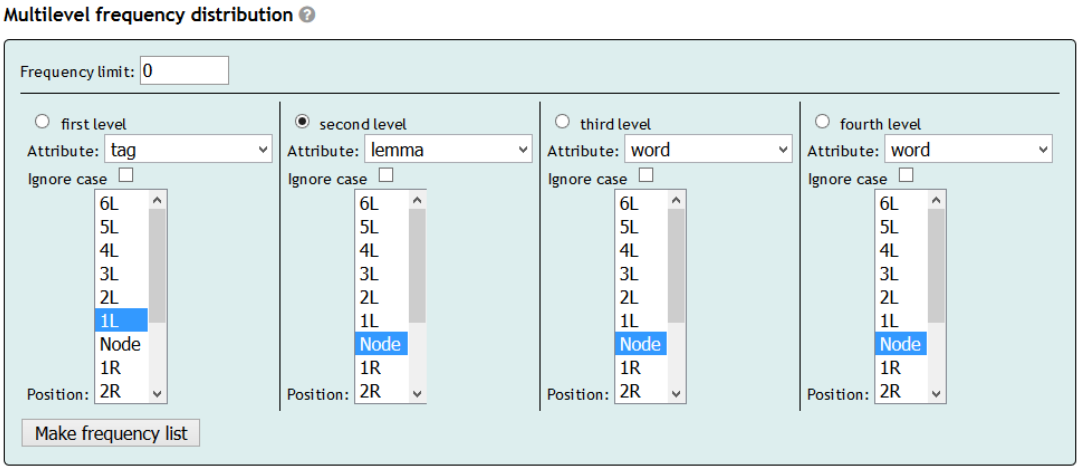


Figure 5-12 Different option for concordance multilevel frequency distribution filter

This is part of the result screen (Figure 5-13), with a combination of tags + lemma *metodologia* ordered by frequency.

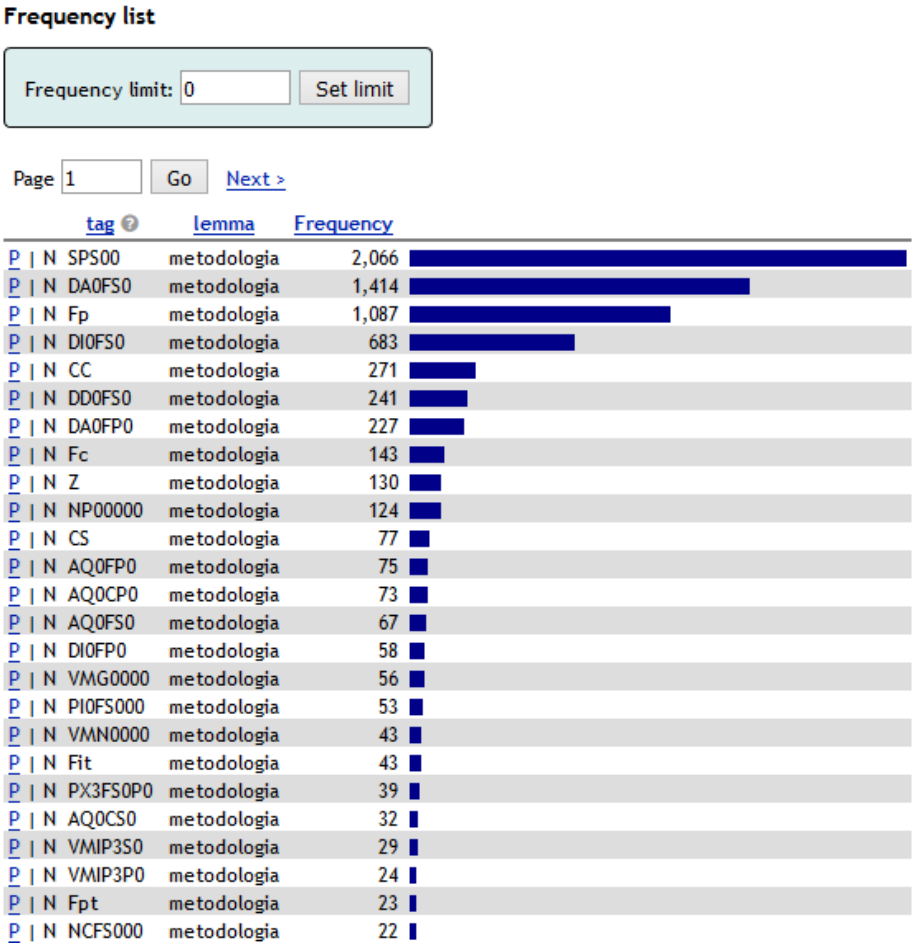
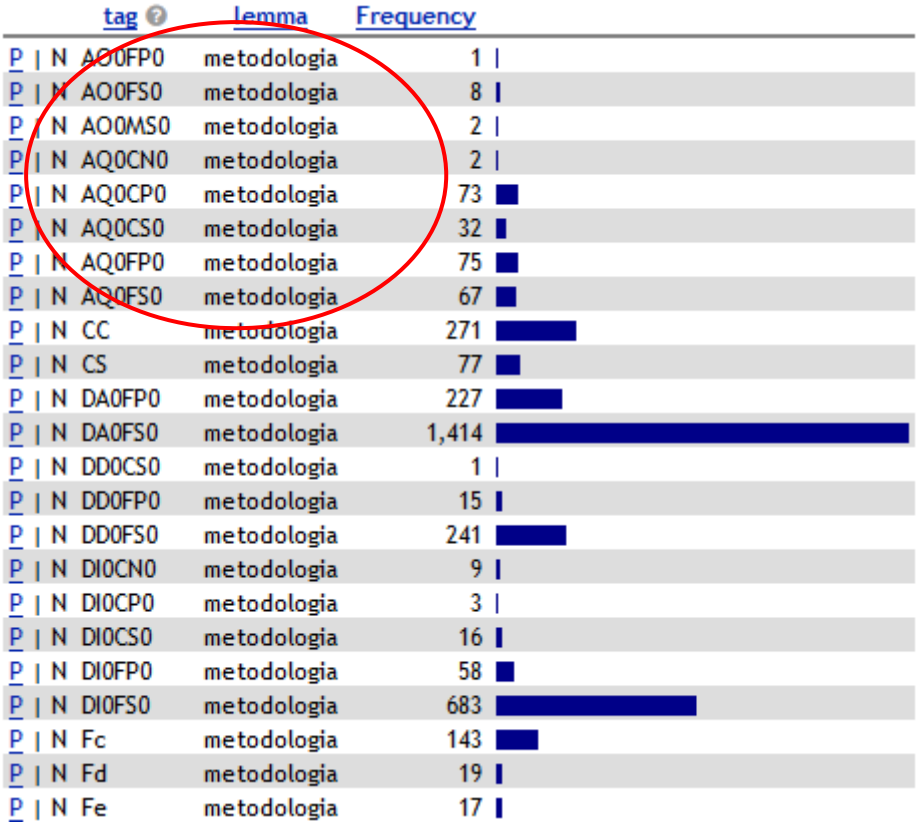


Figure 5-13 Word class of word one to the left of the keyword (*metodologia*) sorted by frequency

When clicked, the question mark in grey next to *tag* opens in a new tab, displaying CoPEP's tagset summary[151] (see Appendix C). Notwithstanding the absence of very interesting additional lexical analyses, it is possible to conclude that adjectives are not among the most frequent word categories preceding *metodologia*. The top five most frequent word classes preceding *metodologia* are preposition (SP), article determiner (DA), period (Fp), indefinite determiner (DI), and coordinating conjunction (CC).

As the focus is on adjectives, which can be represented by A and by VP in the case of participial adjectives, I will sort the results according to the tag name by clicking on *tag* Figure 5-14.

[151] https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-pt.html



tag	lemma	Frequency
P | N AQ0FP0	metodologia	1
P | N AQ0FS0	metodologia	8
P | N AQ0MS0	metodologia	2
P | N AQ0CN0	metodologia	2
P | N AQ0CP0	metodologia	73
P | N AQ0CS0	metodologia	32
P | N AQ0FP0	metodologia	75
P | N AQ0FS0	metodologia	67
P | N CC	metodologia	271
P | N CS	metodologia	77
P | N DA0FP0	metodologia	227
P | N DA0FS0	metodologia	1,414
P | N DD0CS0	metodologia	1
P | N DD0FP0	metodologia	15
P | N DD0FS0	metodologia	241
P | N DI0CN0	metodologia	9
P | N DI0CP0	metodologia	3
P | N DI0CS0	metodologia	16
P | N DI0FP0	metodologia	58
P | N DI0FS0	metodologia	683
P | N Fc	metodologia	143
P | N Fd	metodologia	19
P | N Fe	metodologia	17

Figure 5-14 List of word classes anticipating the keyword, ordered by name tag

As can be seen, there are 260 adjectives in pre-nominal position. By clicking on P on to the left of the tag name, all occurrences are shown. For instance (Figure 5-15):

Query (metodologia)-n 7,397 > Positive filter AQ0FP0 75 > Positive filter metodologia 75 (1.54 per million)

Page 1 of 4 Go Next | Last

1	Exact-Eart...	existentes , além de propostas de *novas* metodologias	/NCFP000/metodologia	, que serão discutidas a seguir . </s><s>
2	Agricultur...	planta . </s><s> Em alternativa , *novas* metodologias	/NCFP000/metodologia	foram introduzidas para a análise genética
3	Agricultur...	, há necessidade de se adequarem *novas* metodologias	/NCFP000/metodologia	para a produção de mudas envasadas em ambiente
4	Health Sci...	não é simplesmente a aplicação de *novas* metodologias	/NCFP000/metodologia	de ensino - aprendizagem . </s><s> A consideração
5	Health Sci...	<s> A adoção , pelas Escolas , de *novas* metodologias	/NCFP000/metodologia	gerou resistências e conflitos devido ao
6	Health Sci...	Problematização , assim como nas *diversas* metodologias	/NCFP000/metodologia	que têm sua fundamentação numa pedagogia
7	Health Sci...	qualificar o corpo docente , adotar *novas* metodologias	/NCFP000/metodologia	de ensino e criar a consciência da necessidade
8	Health Sci...	fotovoz " vem abrindo espaço para *novas* metodologias	/NCFP000/metodologia	e discutindo o quanto a narrativa da visualidade
9	Health Sci...	pesquisas que envolvam a inserção de *novas* metodologias	/NCFP000/metodologia	, ampliando , assim , o fornecimento de
10	Health Sci...	construção coletiva de novos conceitos e *novas* metodologias	/NCFP000/metodologia	que viabilizem sua aplicação . </s><s> Rocha
11	Health Sci...	<s> Sendo importante o emprego de *novas* metodologias	/NCFP000/metodologia	de trabalho envolvendo os pós - graduandos
12	Health Sci...	<s> No entanto , a combinação de *diversas* metodologias	/NCFP000/metodologia	em determinados tipos de estudos pode contribuir
13	Health Sci...	como um problema de saúde . </s> <s> *Novas* metodologias	/NCFP000/metodologia	e abordagens não retiram do trabalho de
14	Applied So...	pesquisadores do tema , utilizando *novas* metodologias	/NCFP000/metodologia	para o enriquecimento dos achados científicos
15	Applied So...	percepção do consumidor pode exigir *novas* metodologias	/NCFP000/metodologia	e uma reavaliação das práticas atuais de
16	Applied So...	Chandra Sekar , em 1949 , com *diversas* metodologias	/NCFP000/metodologia	, como aquelas sugeridas por Lopez e Ruzicka
17	Applied So...	podem ser encontrados avanços nas *diversas* metodologias	/NCFP000/metodologia	de projeções populacionais em estudos como
18	Human Scie...	internacional contemporâneo . </s> <s> *Novas* metodologias	/NCFP000/metodologia	de ação , normas mais cogentes e uma nova
19	Human Scie...	também espaço para discussão de *novas* metodologias	/NCFP000/metodologia	para a elaboração de estatísticas comerciais
20	Human Scie...	serem ministrados e de suas *respectivas* metodologias	/NCFP000/metodologia	requer uma atitude filosófica de reflexão

Figure 5-15 Concordance lines resulting from positive filter

In order to better visualise the results, it is possible to sort the concordances. By clicking on *Sort*, *Left*, the words to the left of the node are ordered alphabetically (Figure 5-15).

Figure 5-16 Left sorting concordance lines

The lexicographer would now take note of the adjectives and their frequency. I will not go further here, though.

The next step is to verify whether participial adjectives are occurring in pre-nominal position (tag VMP) (Figure 5-17):

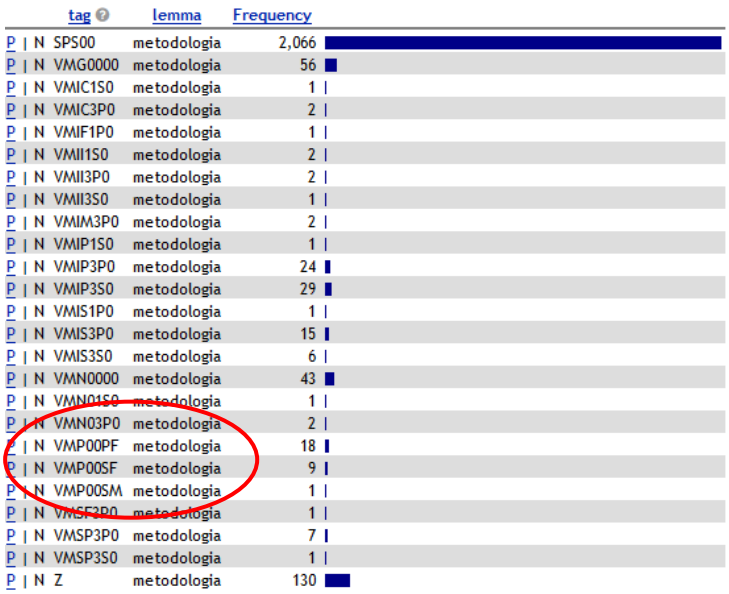


Figure 5-17 Occurrences of participial adjectives in pre-nominal position (keyword=*metodologia)*

There are 28 instances of verbs in participle form. Manual verification of the function of this participial form is required, as it can function as the main verb in a passive voice construction, an auxiliary verb in a compound form or as an adjective. Just as an example, these are some of the results Figure 5-18:

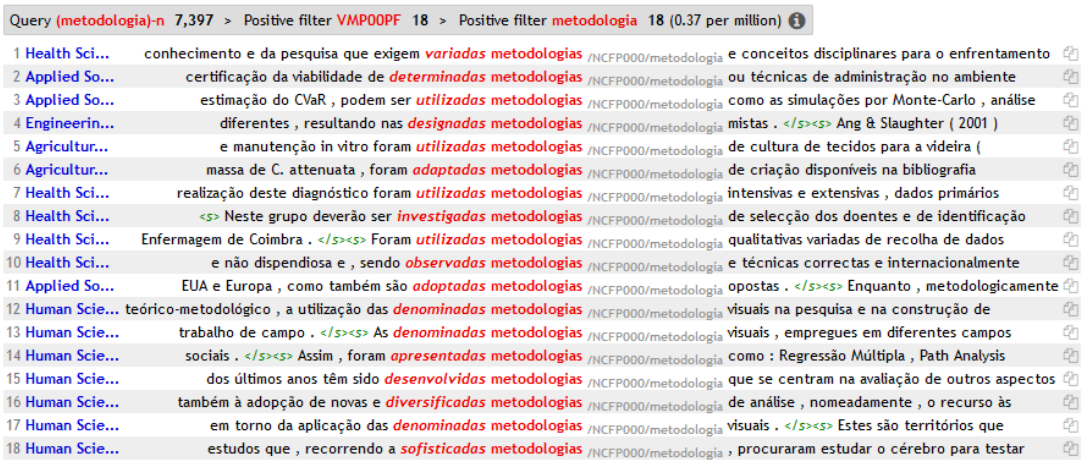


Figure 5-18 Concordances lines for the construction participle forms + *metodologia*

A series of further lexical analyses could be performed. For instance, the lexicographer could repeat the previous routine, only now examining post-nominal adjectives. It would be very interesting to compare adjectives and participial forms used as adjectives in pre- and post-nominal positions.

5.1.1.2 Generating frequency lists with the Search function

The use of CQL significantly enriches query possibilities. One notably beneficial use concerns the generation of very specific frequency lists.

I present here a frequency list of complex prepositions (Shepherd, 2015) formed by three specific elements: preposition+noun+preposition. In order to generate this list, I will type a CQL query in the Search box (Figure 5-19), and click *Make Concordance* (Figure 5-20).

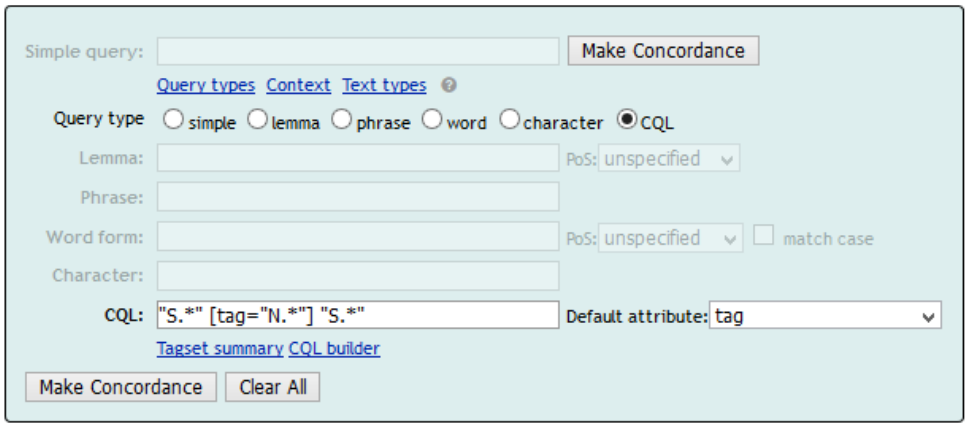


Figure 5-19 CQL search for complex prepositions

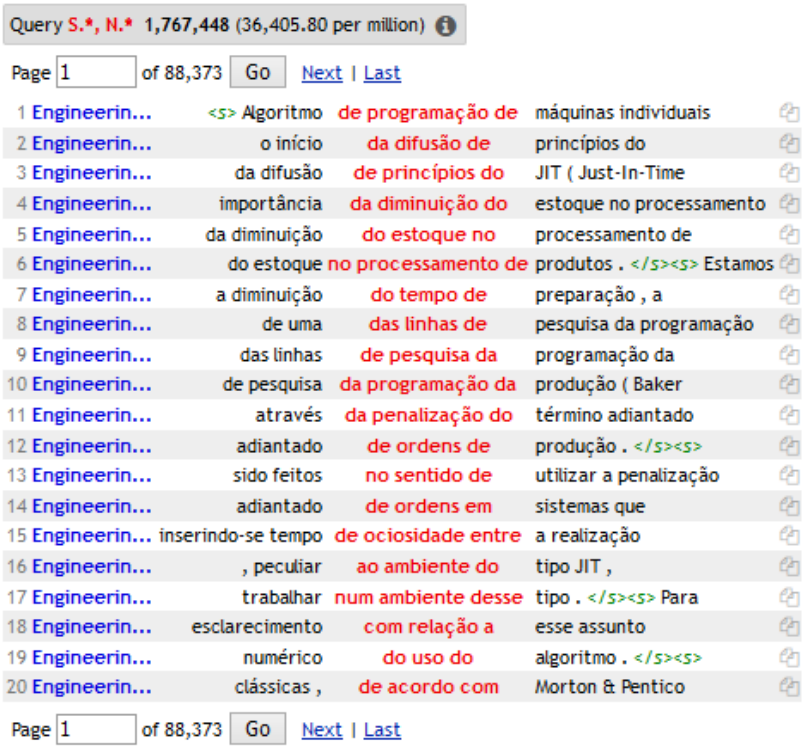


Figure 5-20 Complex prepositions in CoPEP

In order to generate a frequency word list of these constructions, I will sort the results according to *Node form* (option under *Frequency*, on the left menu). This is part of the list (Figure 5-21):

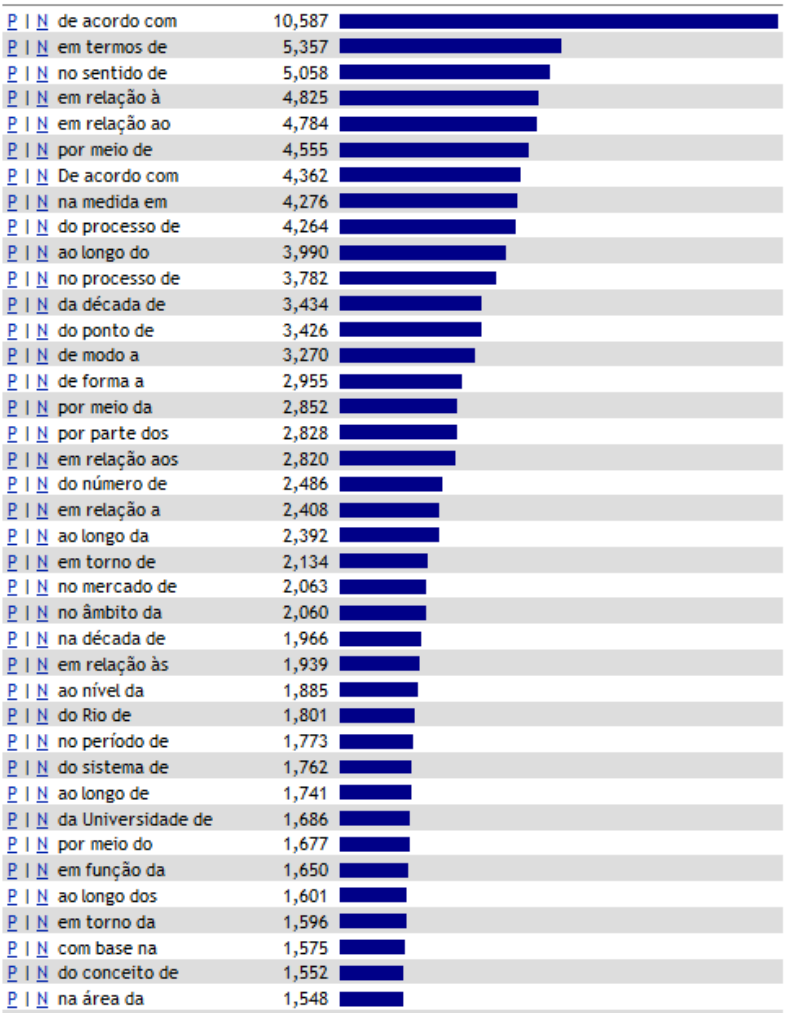


Figure 5-21 Complex prepositions sorted by frequency

This list is a convenient initial point to further carry out lexical analysis concerning multiword expressions, thus contributing to the systematization of lexicographic treatment of such challenging constructions in the dictionary.

There are other useful features for advanced searches that have not been covered here. The Sketch Engine website provides detailed information on them.

5.1.2 Word list

This is the name used to refer to the different frequency lists that can be made in the Sketch Engine, in addition to the most known word list (list of word forms) and lemma list. As with the other functions presented in this chapter, I will not cover all possible lists that can be generated. Instead, some of the lists used in this research will be presented.

5.1.2.1 Lempos list

Lempos stands for lemma with POS-tag indication, as in *estudo*-n ('study'; -n, noun). A list can be created by choosing Lempos in *Search attribute* (Figure 5-22).

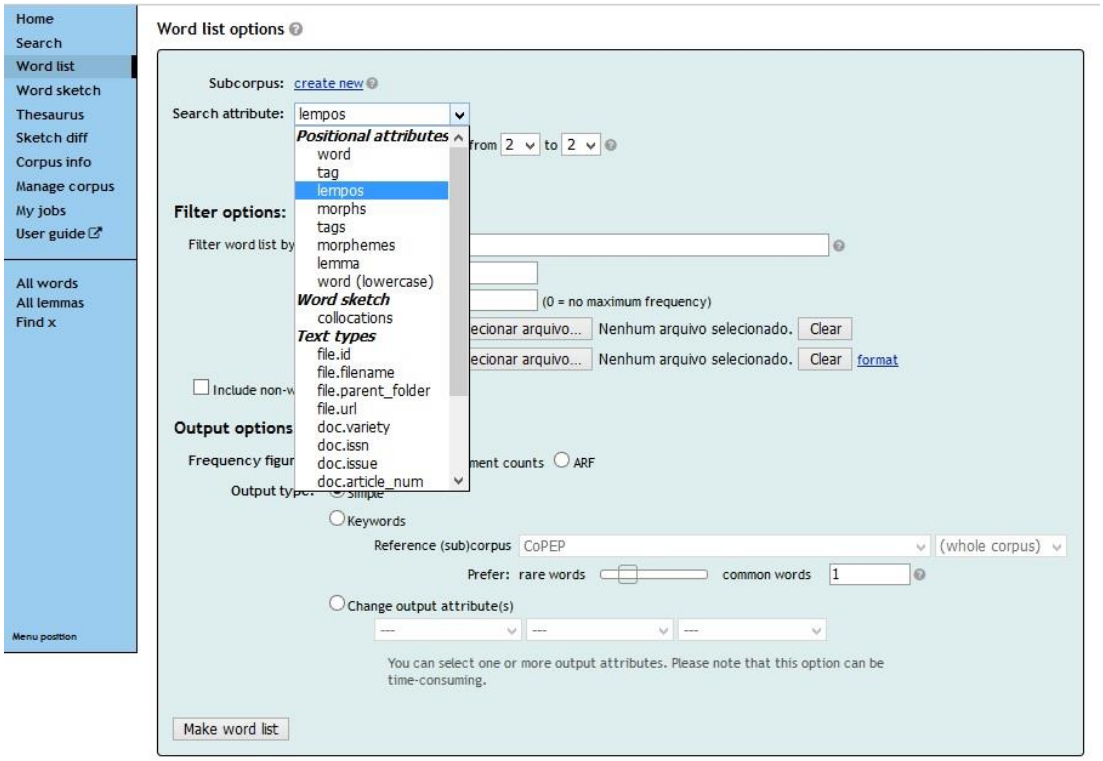


Figure 5-22 Making a lempos list

By clicking *Make word list*, a Lempos list is created. Part of the result can be seen in Figure 5-24.

It is also possible to apply filters in order to restrict the list. For instance, for a list of nouns used in CoPEP, *Lempos* is selected in *Search attribute* and the lempos suffix for noun - -n -, is typed into the *Regular expression* field, under *Filter options* (Figure 5-23). Part of the resulting list is shown in Figure 5-25.

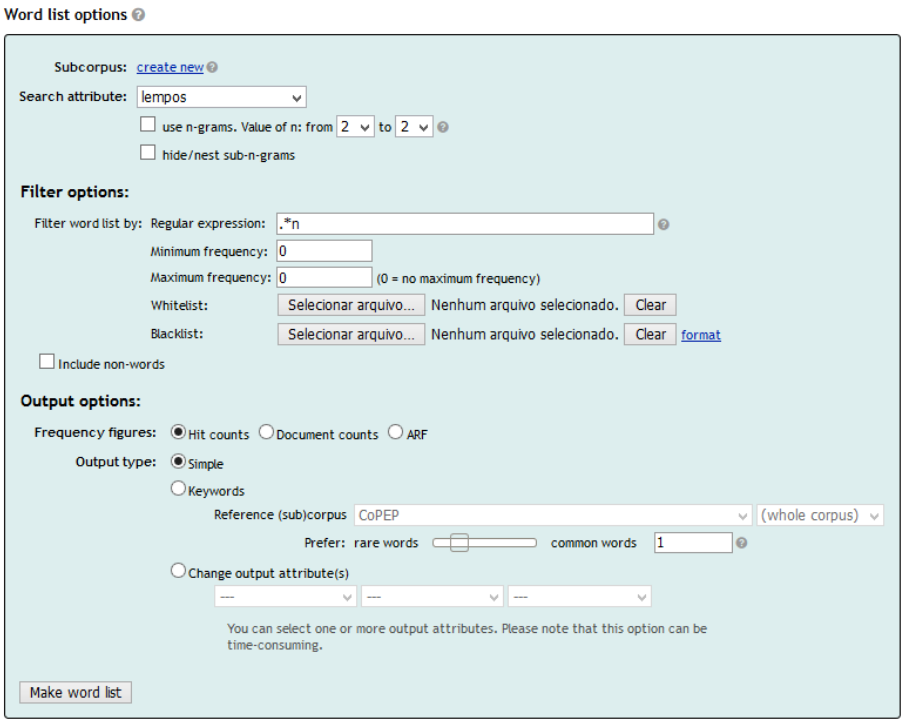


Figure 5-23 Regular expression filter for lempos list making

Word list

Corpus: CoPEP
Total number of items: **93,062**

Page 1 Go Next >

lempos	frequency
de-i	4,281,580
o-x	2,696,051
em-i	1,445,696
e-c	1,276,555
a-i	798,913
ser-v	685,610
um-x	665,467
que-p	595,059
por-i	428,536
para-i	390,269
com-i	384,908
que-c	297,694
não-r	254,420
como-c	244,656
se-p	217,834
ir-v	213,139
ter-v	195,860
mais-r	181,001
ou-c	174,659
poder-v	142,695
o-p	142,503
seu-p	142,002
entre-i	137,767
este-x	105,232
sua-p	104,570
estar-v	97,530
sobre-i	88,098
outro-x	83,090
estudo-n	81,688
também-r	77,313

Figure 5-24 Lempos list

lempos	frequency
estudo-n	81,688
relação-n	76,324
forma-n	71,118
ano-n	70,585
trabalho-n	64,313
saúde-n	60,262
processo-n	59,246
caso-n	51,567
grupo-n	49,931
vez-n	49,690
resultado-n	46,802
análise-n	44,670
nível-n	42,732
país-n	42,213
estado-n	40,959
valor-n	40,885
parte-n	40,366
p-n	39,394
vida-n	39,333
et-n	37,956
sistema-n	37,780
tempo-n	37,693
política-n	36,705
desenvolvimento-n	35,564
al-n	34,491
modelo-n	34,334
pessoa-n	33,964
prática-n	32,827
dado-n	32,745
questão-n	32,045
tipo-n	32,017
área-n	32,010
autor-n	31,334

Figure 5-25 Lempos list- nouns

5.1.2.1.1 N-grams list

It is possible to create lists of n-grams of up to 6 elements by ticking *use n-grams* and defining the values of **n** (Figure 5-26). In *Search attribute*, positional attributes (according to the corpus' original annotation), e.g., word, lemma, and tag, can be chosen to make up the list. In addition, all lists can be further filtered by setting minimum and maximum frequencies.

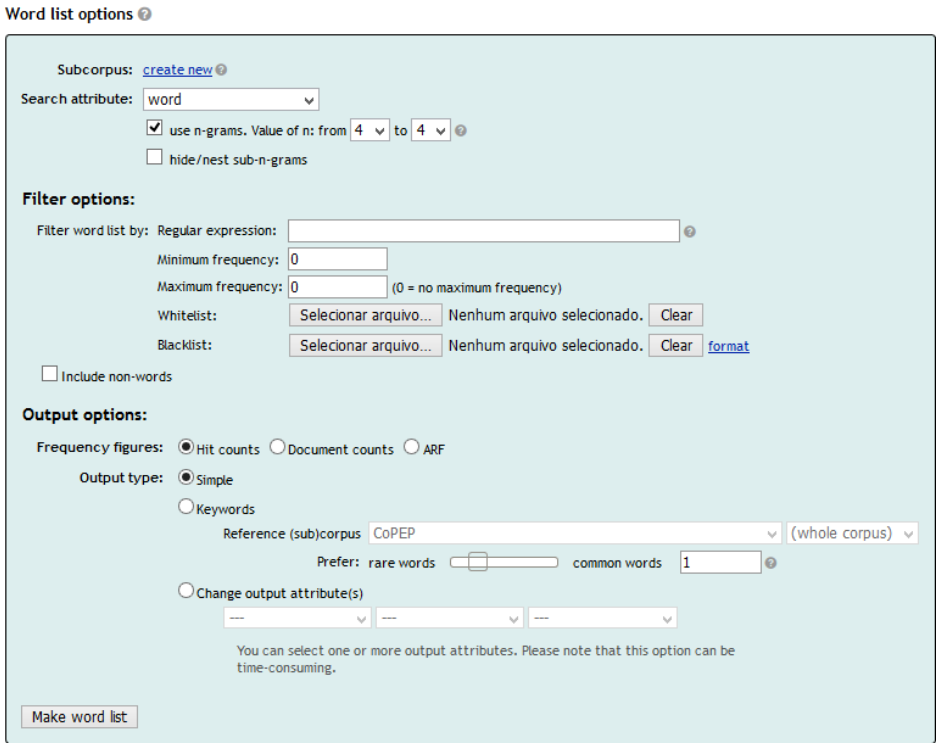


Figure 5-26 Creating a n-gram word list

This n-gram word list creation, where n was set to 4, resulted in a 4-word lexical bundles list, as shown in Figure 5.27.

Another feature of the word list function is generation of a list that either excludes determined attributes (according to the kind of list one is making) via upload of a Blacklist or yields results of only certain pre-defined items, through the upload of a Whitelist. Both Whitelist and Blacklist must be in .txt format.

I want a list of all trigrams in CoPEP, with the exclusion of complex prepositions. For that, I will upload the frequency list of complex prepositions that I created in section 5.1.2 as Blacklist (Figure 5.29). The resulting list can be seen in Figure 5.28.

word (n-grams)	frequency
na medida em que	4,260
do ponto de vista	3,279
de acordo com a	2,793
no que diz respeito	2,442
de acordo com o	2,258
no que se refere	1,979
com o objetivo de	1,963
no mercado de trabalho	1,795
do Rio de Janeiro	1,770
ao longo do tempo	1,746
que diz respeito à	1,724
de acordo com as	1,469
No que diz respeito	1,399
uma vez que a	1,352
que se refere à	1,347
de acordo com os	1,341
o fato de que	1,294
a ideia de que	1,133
como é o caso	1,123
No que se refere	1,088
Do ponto de vista	1,060
a existência de uma	1,018
a partir de uma	1,016
ao mesmo tempo que	1,004
Rio Grande do Sul	1,002
de um conjunto de	994
a qualidade de vida	985
para o desenvolvimento de	981
uma vez que o	955
da qualidade de vida	955
que diz respeito ao	953
com o intuito de	946

Figure 5-27 4-word lexical bundles in CoPEP

word (n-grams)	frequency
uma vez que	9,466
a partir de	8,229
Por outro lado	7,021
ponto de vista	6,451
o processo de	6,221
o número de	5,832
a partir da	5,510
de que a	5,455
a existência de	5,334
um conjunto de	5,147
cada vez mais	5,125
em que a	5,029
Rio de Janeiro	4,785
medida em que	4,751
ao mesmo tempo	4,714
em que o	4,623
qualidade de vida	4,272
de que o	4,218
que diz respeito	4,108
em que se	4,102
a partir do	4,064
de São Paulo	3,966
a presença de	3,777
acordo com a	3,743
mercado de trabalho	3,739
por sua vez	3,585
mais do que	3,515

Figure 5-28 Trigrams in CoPEP excluding complex prepositions

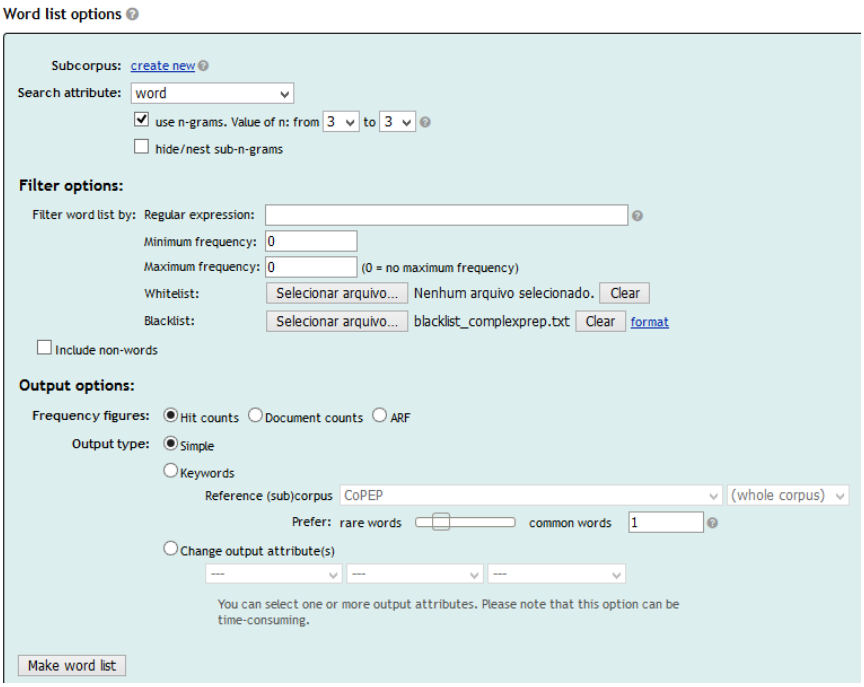


Figure 5-29 Word list creation using Blacklist filter

5.1.2.1.2 Average reduced frequency

Under *Output options*, *Frequency figures* might be *Hit counts*, i.e. number of occurrences of the search attribute in the corpus or subcorpus at play, *Document counts*, i.e. the number of documents (in the case of CoPEP, texts) in which the search attribute occurs, and *ARF*, average reduced frequency, which "is a variant on a frequency list that 'discounts' multiple occurrences of a word that occur close to each other, e.g. in the same document".[152] The ARF thus provides an estimate of the frequency of a word if the corpus were homogeneous.[153]

In order to illustrate the ARF result, I will repeat the routine for generation of a lempos frequency list of nouns in CoPEP (section 5.2.1); however, this time the output frequency figure will not be hit counts but ARF (Figure 5-30).

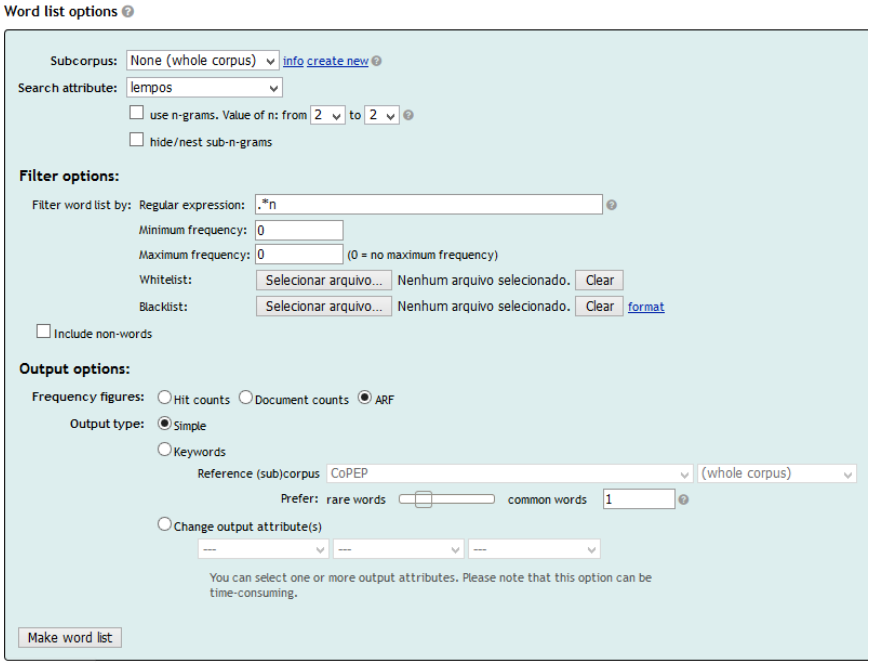


Figure 5-30 Creating a lempos frequency list with ARF output

I present here the two noun frequency lists in CoPEP. In Figure 5-31, a partial hit counts list is shown, while in Figure 5-32, the list is composed of ARF values.

[152] https://www.sketchengine.co.uk/documentation/average-reduced-frequency/

[153] I suggest the documentation on ARF available on the Sketch Engine website for a very elucidative explanation of how the calculation works. https://www.sketchengine.co.uk/documentation/average-reduced-frequency/

lempos	frequency
estudo-n	81,688
relação-n	76,324
forma-n	71,118
ano-n	70,585
trabalho-n	64,313
saúde-n	60,262
processo-n	59,246
caso-n	51,567
grupo-n	49,931
vez-n	49,690
resultado-n	46,802
análise-n	44,670
nível-n	42,732
país-n	42,213
estado-n	40,959
valor-n	40,885
parte-n	40,366
p-n	39,394
vida-n	39,333
et-n	37,956
sistema-n	37,780
tempo-n	37,693
política-n	36,705
desenvolvimento-n	35,564
al-n	34,491
modelo-n	34,334
pessoa-n	33,964
prática-n	32,827
dado-n	32,745
questão-n	32,045
tipo-n	32,017
área-n	32,010
autor-n	31,334

Figure 5-31 Lemma list-frequency ordered

lempos	average reduced frequency
forma-n	38,367.70
relação-n	37,484.00
estudo-n	33,219.80
ano-n	30,065.60
vez-n	27,857.60
processo-n	25,982.70
trabalho-n	25,364.50
caso-n	23,930.90
parte-n	22,069.60
resultado-n	20,575.00
análise-n	19,463.20
tempo-n	18,633.20
grupo-n	18,094.10
nível-n	17,694.10
vida-n	15,541.20
estado-n	15,398.70
valor-n	15,322.30
tipo-n	15,221.70
questão-n	15,003.70
desenvolvimento-n	14,849.20
a-n	14,701.70
sentido-n	14,464.00
sistema-n	14,364.60
país-n	14,253.80
modo-n	14,104.50
saúde-n	13,530.50
autor-n	13,275.80
contexto-n	12,906.00
condição-n	12,894.30
problema-n	12,877.10
exemplo-n	12,873.20
dado-n	12,832.40
prática-n	12,774.90
situação-n	12,669.20
necessidade-n	12,443.70
pessoa-n	12,327.40
área-n	12,276.20

Figure 5-32 Lemma list - ARF ordered

5.1.2.2 Comparing corpora

Generally used to determine specialized terms related to a certain corpus topic, the *Keywords* output option can also be used to compare two corpora (or subcorpora) in order to generate a frequency list of a corpus' typical words. Under *Output type*, a *Reference (sub)corpus* is chosen to be the reference against which the focus corpus will be compared. In *Prefer*, below the Reference corpus name, a slider allows generating a term-oriented list if the *rare words*-end is selected, while *common words*-end results in a comparison of frequent words between the two corpora (Figure 5-34).

In order to illustrate this function, I will compare the subcorpora of the two language varieties in CoPEP to see which common words are used more in one variety than the other. I will begin with the European Portuguese (EP) subcorpus as the focus and the Brazilian Portuguese (BP) subcorpus as the reference (Figure 5-33). The results are displayed in Figure 5-35.

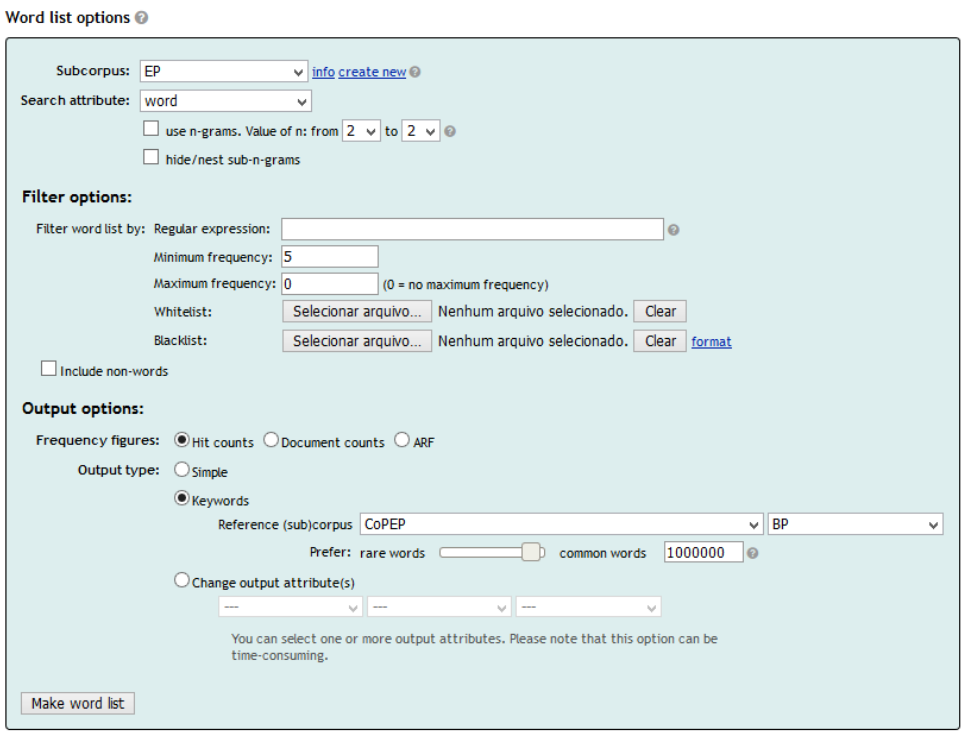


Figure 5-34 Using keyword output option for comparing (sub)corpora

Corpus: CoPEP
Subcorpus: EP

Reference corpus: CoPEP
Reference subcorpus: BP
Switch focus and reference (sub)corpus

Page 1 Go Next >

	CoPEP : EP		*CoPEP : BP*		
word	frequency	frequency/mill	frequency	frequency/mill	Score
e	652,514	26916.7	612,184	25186.0	1.0
a	676,017	27886.2	645,673	26563.8	1.0
com	199,900	8246.0	172,682	7104.4	1.0
uma	173,420	7153.7	155,678	6404.8	1.0
Portugal	18,101	746.7	1,392	57.3	1.0
os	195,508	8064.8	180,620	7430.9	1.0
dos	165,833	6840.7	151,247	6222.5	1.0
doentes	15,841	653.5	1,184	48.7	1.0
facto	13,595	560.8	113	4.6	1.0
et	24,852	1025.2	12,940	532.4	1.0
estudo	31,073	1281.8	19,567	805.0	1.0
al	22,714	937.0	11,240	462.4	1.0
esta	21,057	868.6	9,934	408.7	1.0
numa	15,715	648.3	4,744	195.2	1.0
mais	93,962	3876.0	83,466	3433.9	1.0
num	15,533	640.7	4,873	200.5	1.0
nível	19,414	800.8	9,275	381.6	1.0
este	21,129	871.6	11,343	466.7	1.0
investigação	11,742	484.4	2,471	101.7	1.0
quer	10,931	450.9	2,624	108.0	1.0
através	16,568	683.4	8,319	342.3	1.0
nomeadamente	7,998	329.9	114	4.7	1.0
crianças	11,941	492.6	4,299	176.9	1.0
resultados	20,940	863.8	13,603	559.6	1.0
as	148,769	6136.8	141,789	5833.4	1.0
ou	88,812	3663.6	81,819	3366.1	1.0
das	116,827	4819.2	109,975	4524.5	1.0
the	12,379	510.6	5,597	230.3	1.0
controlo	6,648	274.2	15	0.6	1.0

Figure 5-33 Comparison between EP corpus (focus) and BP (reference).

Reference corpus: CoPEP
Reference subcorpus: EP
Switch focus and reference (sub)corpus

Page 1 Go Next >

	CoPEP : BP		CoPEP : EP		
word	frequency	frequency/mill	frequency	frequency/mill	Score
do	342,367	14085.4	302,741	12488.3	1.0
o	415,162	17080.3	381,362	15731.4	1.0
de	1,131,123	46535.8	1,095,129	45174.8	1.0
em	271,093	11153.1	238,899	9854.8	1.0
Brasil	25,513	1049.6	4,111	169.6	1.0
para	190,262	7827.6	171,973	7094.0	1.0
no	159,563	6564.6	141,685	5844.6	1.0
enfermagem	16,352	672.7	2,381	98.2	1.0
como	137,653	5663.2	123,401	5090.4	1.0
fato	13,309	547.5	954	39.4	1.0
não	125,634	5168.7	114,054	4704.8	1.0
ele	15,503	637.8	5,198	214.4	1.0
se	120,516	4958.2	110,157	4544.1	1.0
saúde	29,012	1193.6	19,110	788.3	1.0
é	136,722	5624.9	126,680	5225.6	1.0
meio	15,283	628.8	5,656	233.3	1.0
trabalho	32,143	1322.4	22,498	928.1	1.0
essa	15,808	650.4	6,474	267.1	1.0
governo	11,819	486.2	2,594	107.0	1.0
política	23,808	979.5	14,841	612.2	1.0
esse	14,292	588.0	5,488	226.4	1.0
pesquisa	13,405	551.5	4,978	205.3	1.0
pacientes	10,899	448.4	2,713	111.9	1.0
Enfermagem	9,368	385.4	1,489	61.4	1.0
controle	8,672	356.8	942	38.9	1.0
processo	25,478	1048.2	17,745	732.0	1.0
sobre	46,954	1931.7	39,416	1625.9	1.0
ações	9,821	404.0	2,408	99.3	1.0
cuidado	9,197	378.4	1,857	76.6	1.0
brasileira	8,311	341.9	977	40.3	1.0
São	11,034	454.0	3,944	162.7	1.0
Paulo	8,512	350.2	1,469	60.6	1.0

Figure 5-35 Comparison between BP corpus (focus) and EP (reference)

5.1.3 Word Sketch

The word sketch is the hallmark of the Sketch Engine and is undoubtedly the most sophisticated corpus query function currently available. It provides a lexical profile summary of a word in one result screen, gathering in one place a word's most typical grammatical and collocational behaviour.

A general overview of the word sketch was given in Chapter 3, while the history of its development and the turning-point in lexicographic work it has led to are presented in 3.3. Here I will explore some other features of the word sketch.

The main statistics behind word sketches is logDice, while since 2015 other statistics have been added to provide better scores for specific computations. It is beyond the scope of this thesis to describe statistical calculations. I refer to the documentation available on the Sketch Engine website[154] for an elucidative presentation of all statistics used in the corpus tool.

[154] https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/

The simplest way to generate a word sketch is by typing into the *Lemma* field a lemma and either leave the system to display the different word categories of the lemma by selecting *auto* in the drop-down box in *Part of speech* (Figure 5-36) or define which word class is wanted (Figure 5-37).

Figure 5-36 Generating a word sketch search. POS: auto

Figure 5-37 Generating a word sketch search. Defined POS

A partial view of the results of the first kind of word sketch generation for *político* ('political', adjective; 'politician', noun), i.e. with *auto* selected, can be seen in Figure 5-38.

político *(adjective)* Alternative PoS: noun (1,280) adverb (3)
CoPEP freq = 46,384 (955.41 per million)

usage patterns		
+infinitivo	81	0.17

N mod por político_Adj-Part		
	38,309	82.59
partido +	1,216	9.91
dos partidos políticos		
sistema +	1,309	9.56
do sistema político		
poder +	854	9.25
do poder político		
regime +	500	8.56
regime político		
instituição +	505	8.51
das instituições políticas		
decisão +	458	8.46
decisões políticas		
participação +	437	8.41
participação política		

e_ou		
	2,579	5.56
económico +	356	11.80
política e económica		
econômica +	176	10.97
política e econômica		
econômicas +	146	10.74
políticas e econômicas		
ideológico +	144	10.66
político e ideológico		
econômico +	121	10.47
político e econômico		
administrativo +	116	10.35
política e administrativa		
econômicos +	110	10.34
, políticos e econômicos		

Figure 5-38 Partial word sketch for *político*, "auto"

In the header, the lemma (*político*) is displayed with the word category between brackets (adjective). Since *auto* had been selected, other occurrences of the lemma tagged with a different part of speech are shown with frequency in parentheses; in the example here, *politíco* as a noun, with 1,280 occurrences, and as an adverb, with three occurrences. The corpus name is shown (CoPEP), and frequency of the lemma as an

adjective in this corpus can be seen (46,384), together with its normalized frequency per million (955.41).

The headings of each column indicate the grammatical relation (gramrel) in which the lemma participates (here, noun modified by an adjective and symmetric *and/or* structure), the total frequency of that construction and the overall score, i.e. the collocation strength. By clicking on the frequency number, all concordances are displayed, while a click on the name of the gramrel opens up a new tab with the sketch grammar (more information below).

Below the heading, a list of collocates is displayed with raw frequencies and scores, which are used to order the collocates. In the present example (Figure 5-39), collocates are ordered by score.

Let us take the gramrel N mod por Adj-Part for further exemplification (Figure 5-39). As can be seen, the top collocate for *político* is *partido* ('party'). The grey phrase in the line beneath the collocate is called Longest-Commonest Match (LCM) and is the most frequent combination found in the corpus. The LCM function provides a quick understanding of the collocation. In this example, the *multi-word sketch function* is also available. By clicking on the plus sign by the collocate in bold, another word sketch is generated, now showing collocates for the collocation in question.

Figure 5-39 Partial view for grammatical relation noun modified by adjective, lemma *político*

I will first see the concordances of the collocation *político+partido*. For that, I will click on the frequency number (1,216) next to the collocate *partido*. This is the screen result (Figure 5-40):

Figure 5-40 Concordance lines for *partido* (noun) *político* (adjective)

As can be seen, this is the regular concordance pane, with all functionalities that have been previously demonstrated (see Figure 5-5 in section 5.1.1), meaning that further refinement of the analysis can be performed. Concordances can be sorted (e.g., right, left, node), sampled (a random number of lines defined by the user is picked by the system), and sorted by frequency, among other functions.

When clicking on the plus sign next to a collocate, a new word sketch is generated for the collocation in question. For instance, our example *partido político* systematically combines with other words, as can be seen in the multi-word sketch result screen in Figure 5-41, after clicking on + next to *partido* (as shown in Figure 5-39 above).

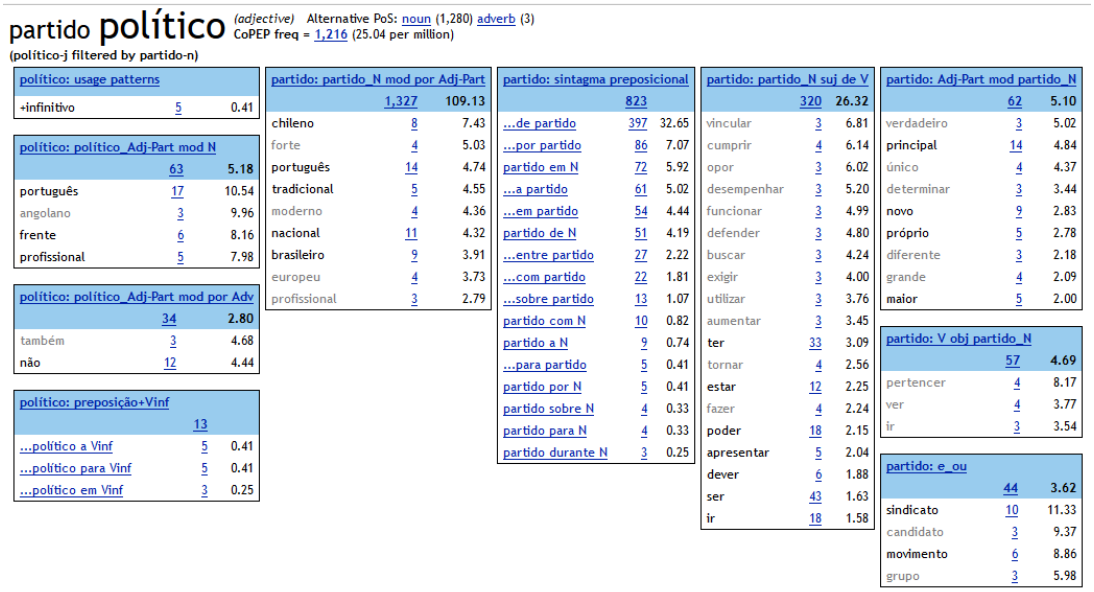


Figure 5-41 Multi word sketch *partido político* +

Let us have two close-ups for further illustration. Figure 5-42 shows the gramrel adjective modifying *partido*, while Figure 5-43 displays the concordances after clicking the collocate *principal* and revealing the collocation *principal partido político* ('main political party').

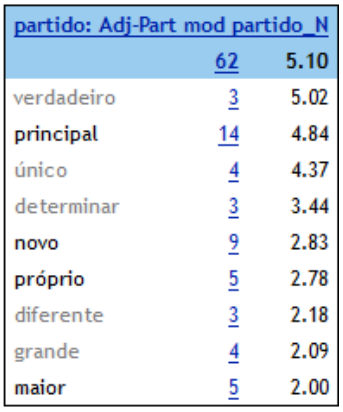


Figure 5-42 Word sketch for gramrel adjective modifying *partido*

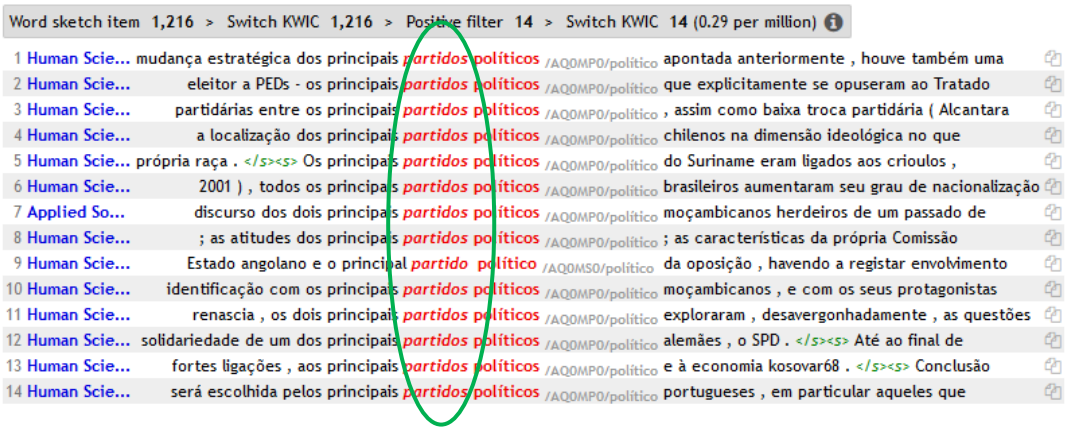


Figure 5-43 Concordances of a multi word sketch, *principal partido político*

On the result word sketch screen (Figure 5-44), in the lower half of the left menu, there are some additional features. In *Save*, all that is displayed on the page can be saved as txt or xml. *Change options* returns to the initial page. *Cluster* is a function for grouping collocates according to their meaning similarity, which is defined based on similarity of collocational behaviour[155] (information on clustering values for CoPEP in Chapter 9). *Sort by freq* toggles between frequency and scores for determining collocate order. *Hide gramrels* displays the results as a continuous list rather than in boxes. *More* or *Less data* allows expanding or reducing the amount of data displayed.

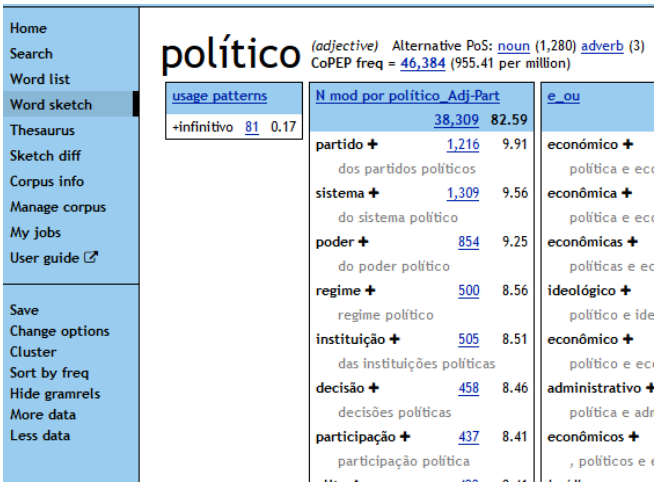


Figure 5-44Additional options on lower-half part of the left menu

[155] For information on statistics behind the cluster function, see https://www.sketchengine.co.uk/documentation/clustering-neighbours-documentation/

5.1.3.1 Advanced options – Tickbox Lexicography

More experienced users can benefit from advanced options that permit customizations of word sketch generation and the displaying of results. Figure 5-45 depicts what is seen when *Advanced options* is clicked on the initial page. Nevertheless, due to space constraints, attention will be drawn to Sketch Engine-exclusive Tickbox Lexicography (TBL) (Granger & Paquot, 2010),[156] which is a highly valuable function for lexicographers.

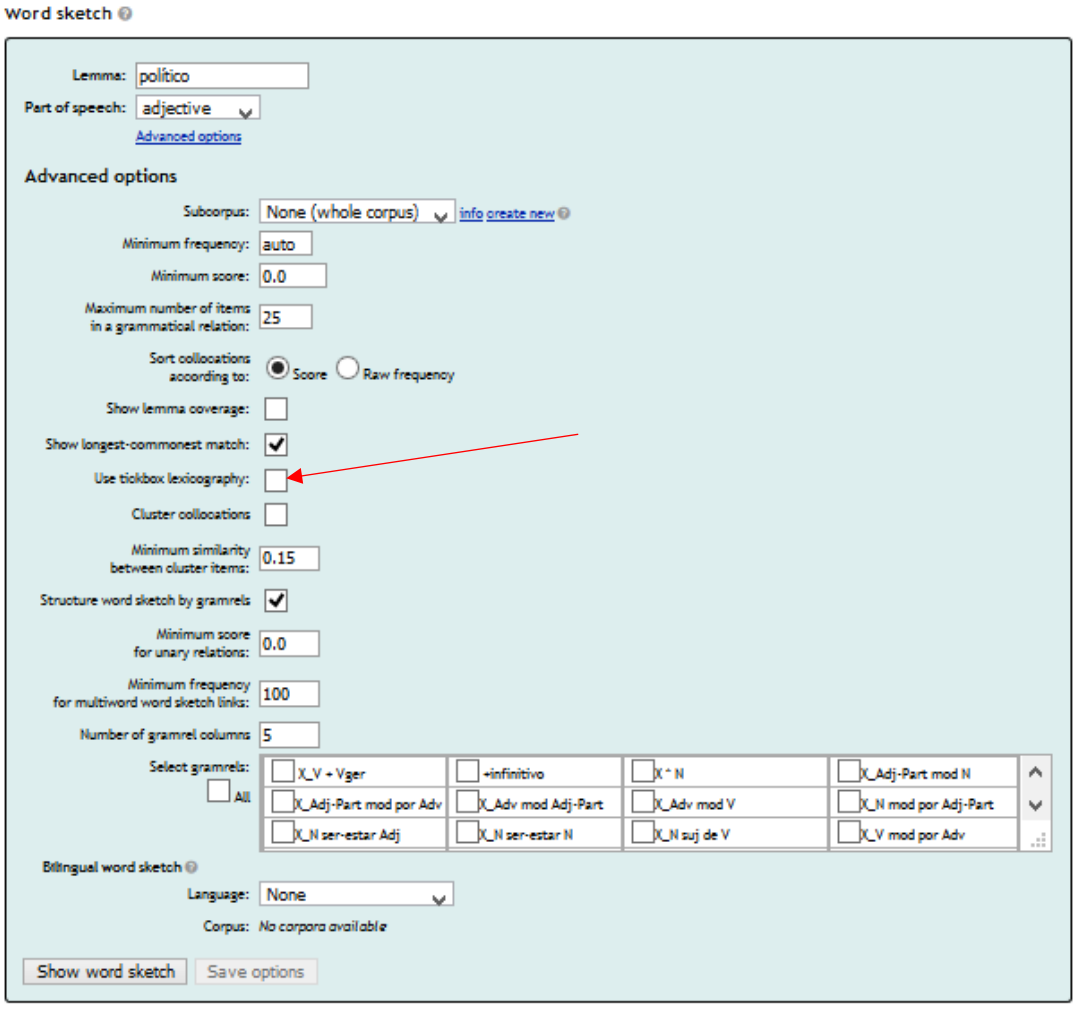


Figure 5-45 Advanced options in word sketch

Tickbox Lexicography (TBL) is a powerful tool that streamlines lexicographic work[157] by enabling users with a direct selection of examples from the corpus and

[156] TBL is not available for users by default, requiring additional payment of an annual fee.

[157] It can also be used by teachers and professional working on teaching materials.

import into the dictionary writing system. Ticking off *Use tickbox lexicography* (TBL) activates this function. Since TBL depends on the Good Dictionary Examples function (GDEX) (Kilgarriff et al., 2008), let us first briefly touch upon this feature before presenting TBL.

GDEX is a function that automatically analyses the corpus for good examples based on pre-defined criteria, scores the concordances and provides lexicographers with the best examples at the top of the list. Thus, by evaluating the top sentences, the whole fastidious work of analysing hundreds of examples is reduced to only a few (defined by the user).

For GDEX to know what good examples are, it must be informed by lexicographers with a number of classifiers (e.g., sentence length, frequency of words) that make up a determined configuration. That means that for each language, including possible further refinement for different kinds of dictionary users' needs, a GDEX configuration must be devised. GDEX configurations are uploaded to the Sketch Engine by clicking on *GDEX configuration*, under *Advanced Features* on the initial page of the Sketch Engine, as shown in Figure 5-46.

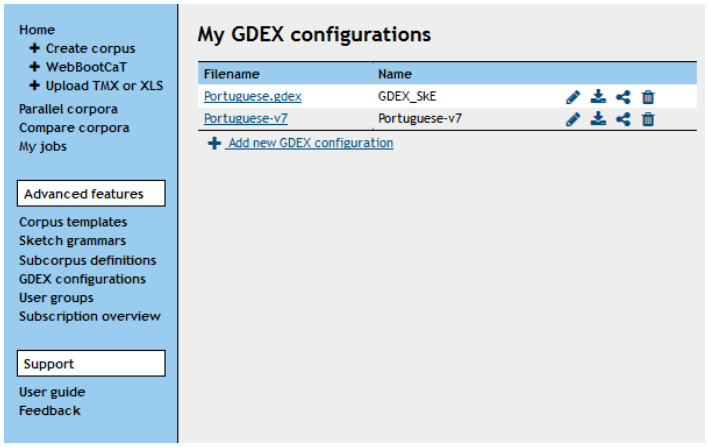


Figure 5-46 Advanced features, uploading a GDEX configuration

I will not go into further details here as Chapter 8 presents the special GDEX configuration that I have developed for DOPU using CoPEP.

Moving back to TBL, this function works in the following way. A word sketch is generated for a certain lemma (here, *político* -adjective). The result screen is displayed with boxes next to each collocate (Figure 5-47).

Figure 5-47 Word sketch with activated TBL

The lexicographer then ticks the boxes next to the chosen collocates and is taken to a page with the top examples (the number of examples is defined by the user) sorted according to scores computed by the system according to a certain GDEX configuration (Figure 5-48).

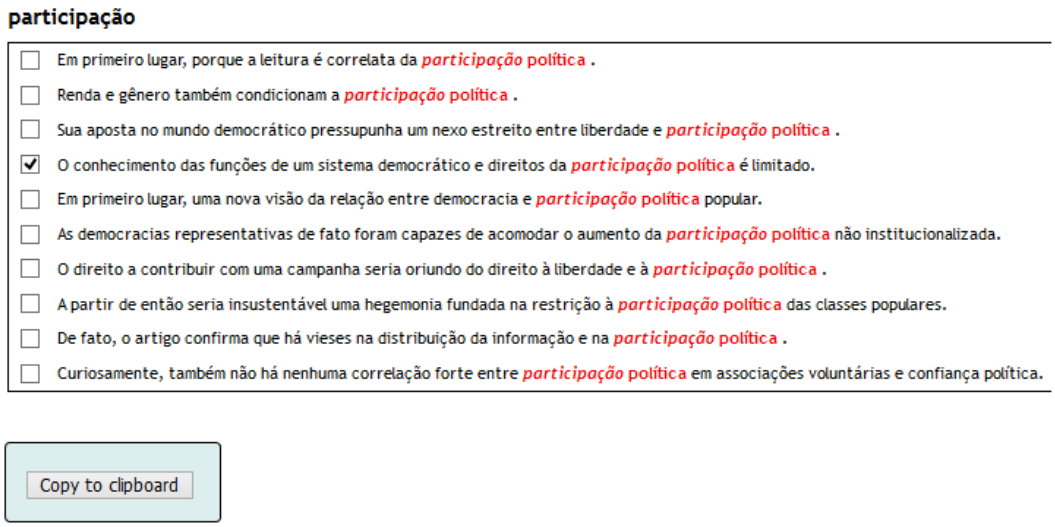


Figure 5-48 TBL with GDEX

The example is saved in an XML file (Figure 5-49), which can be customised to match the DTD of the dictionary writing system.

```
<entry>
<keyword>político</keyword>
<gramrel>
  <grname>N mod por X_Adj-Part</grname>
  <collocation>
    <collo>participação</collo>
    <example>O conhecimento das funções de um sistema democrático e direitos da
    participação <b>política</b> é limitado.</example>
  </collocation>
</gramrel>
</entry>
```

Figure 5-49 XML file of example selected with TBL

5.1.3.2 Sketch Grammar

The condition for generation of word sketches and word sketches difference (see 5.4 below) is a sketch grammar. This is a file with grammatical relations comprising CQL-written queries that find collocations via pattern-matching over a corpus. Sketch grammar is language dependent and can be customised to attend to specific users' needs. Sketch grammars are devised and uploaded to the Sketch Engine under Advanced features (Figure 5-46).

Figure 5-50 shows a partial view of AcadPortSkG. I will not go further into the subject here as Chapter 6 describes in detail the process of devising the AcadPortSkG, the sketch grammar that I have created especially for CoPEP.

```
*DUAL
=%w_N mod por Adj-Part/N mod por %w_Adj-Part
1:"N.*""R.*"{0,2}"A.*|V.P.*"{0,2}2:"A.*|V.P.*"

*DUAL
=%w_N ser-estar N/N ser-estar %w_N
1:"N.*" "R.*"{0,2} [lemma="ser|estar"] "A.*|V.P.*|D.*|Z.*|P.*"{0,4} 2:"N.*"
```

Figure 5-50 Partial view of sketch grammar AcadPortSkG

5.1.4 Thesaurus and Sketch Diff

This is not a traditional thesaurus in the common sense of a dictionary of synonyms. Instead, it is a distributional thesaurus that uses special algorithms[158] for

[158] For a detailed description of the computation of Thesaurus, see https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/

finding words that share similar grammatical behaviour contexts. The quality of the results depends on the size of the corpus: the larger, the better.

For generation of a Thesaurus result for a lemma, it is possible to follow a simple routine, with lemma typing into *Lemma field* and selection of POS attribute in a drop-down box. Advanced options include results clustering and Maximum number of items, as can be seen, in Figure 5-51. After hitting Show similar words, the screen in Figure 5-52 is displayed.

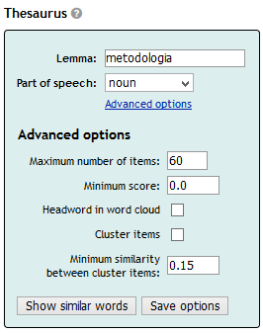


Figure 5-51 Thesaurus search with display of advanced options

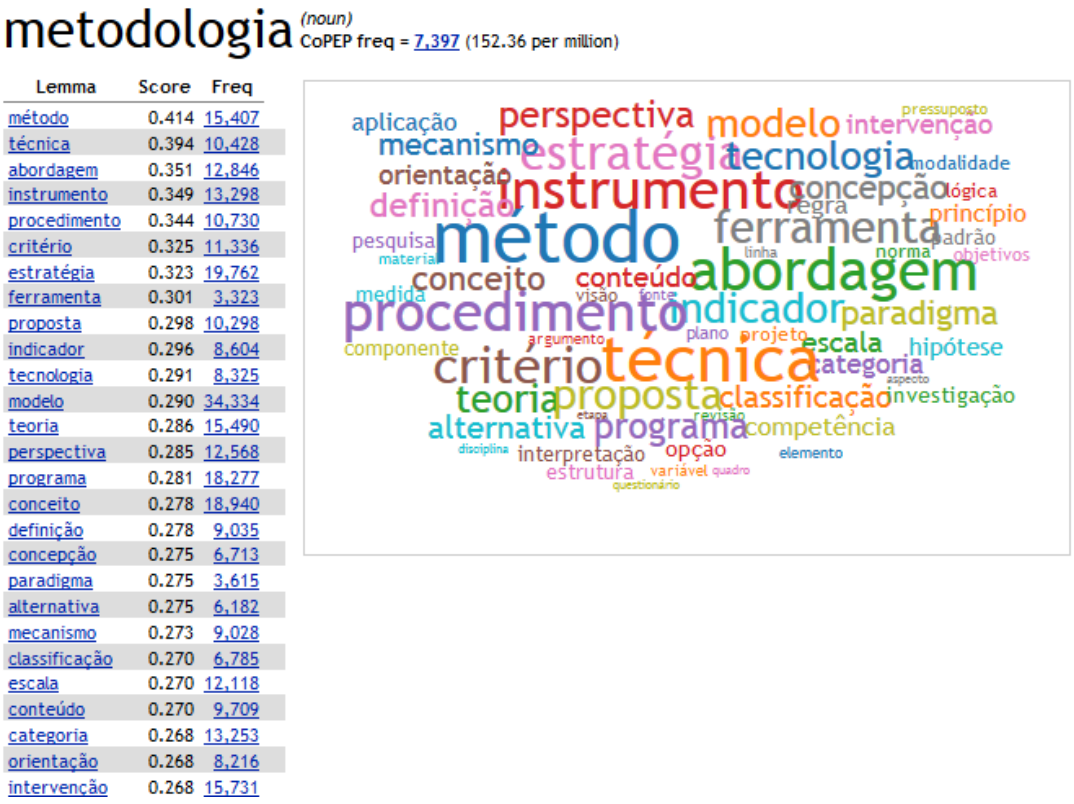


Figure 5-52 Thesaurus search result with list and word cloud visualization

In the header, the search lemma, its word class, the corpus used, frequency and frequency per million of the lemma are shown. On the left-hand side of the pane, a list of similar lemmas is shown, sorted by score and with frequency of occurrence in CoPEP, while in the centre the results are visualised in a cloud. All synonyms are clickable. When a lemma is clicked, a word sketch difference result screen opens up, as in Figure 5-53. It is also possible to access word sketch difference function by directly clicking on Sketch Diff, in the upper-half of the left menu on the initial page.

5.1.4.1 Sketch Diff

Word sketches difference compares the collocational behaviour of two lemmas, thus helping, for example, to disambiguate senses. Above, in Thesaurus, I looked for synonyms for the lemma *metodologia* (‘methodology’) The top word is *método* (‘method’). By clicking on it (either on the list to the left or inside the cloud), this is the screen result that shows up Figure 5-53:

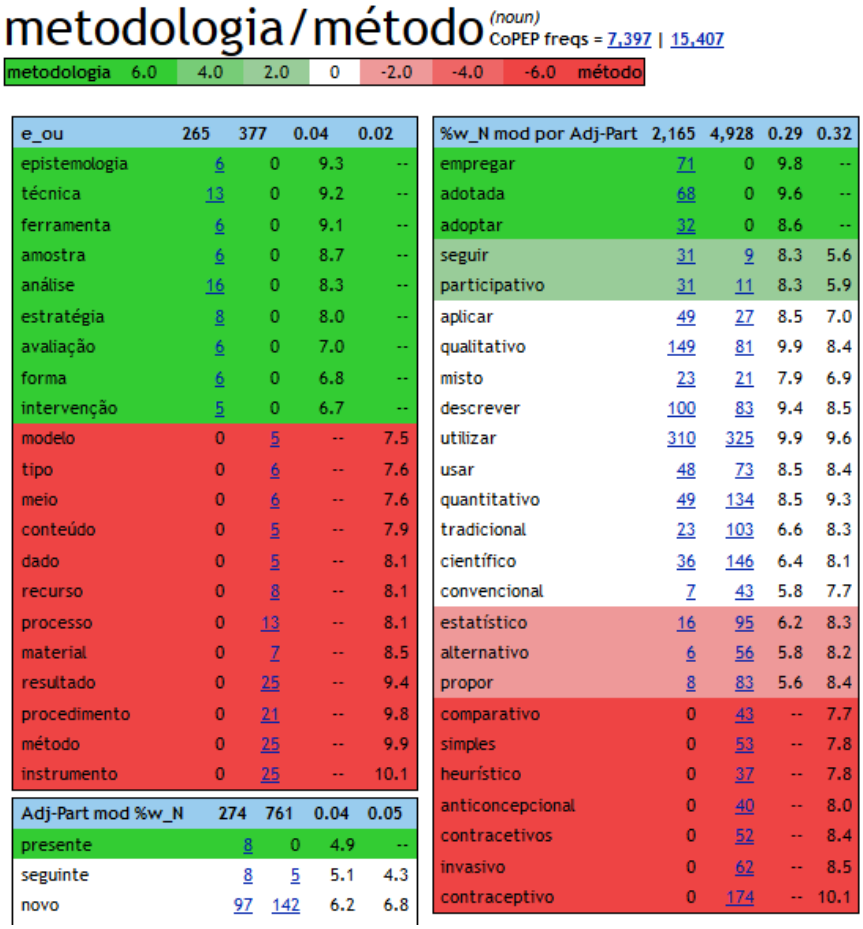


Figure 5-53 Sketch Diff result screen

In the header, there are two frequencies, in reference to the first and second lemmas respectively. The colour bar beneath the header is composed of three colours: green to the left end, white in the middle and red to the right end. Extreme-end green indicates the collocate occurs only with the first lemma. Extreme-end red indicates the collocate occurs only with the second lemma. The white zone means that the collocate occurs with both lemmas. In between white and the other two colours, there is a gradient of both green and red in reference to areas of some overlapping of occurrences.

Headings of each box (in blue) show the grammatical relation, while the first column shows the frequency count for the collocate with the first lemma and the second column, the frequency of co-occurrences of the same collocate with the second lemma. Column 3 and 4 are used for development purposes only. Frequencies are clickable and go to KWIC concordances.

Another feature in word sketch difference is the possibility to compare two word forms of the same lemma (see Figure 5-54). For instance, in Portuguese, the conjugation of the first person present indicative of the verb *ouvir* (‘hear’) has two forms: *ouço* and *oiço*.

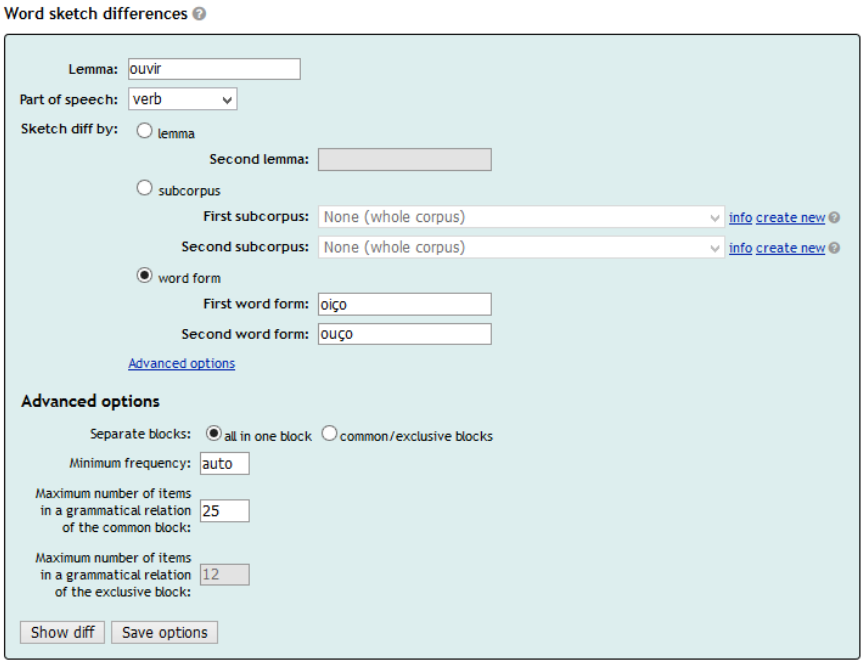


Figure 5-54 Word sketch differences search (*oiço*, *ouço*)

I will use the 2.8-billion-word Portuguese Web 2011 (PtTenTen 11, Palavras parsed) corpus, which covers EP and BP varieties of texts crawled from the Internet, to demonstrate the comparison (Figure 5-55).

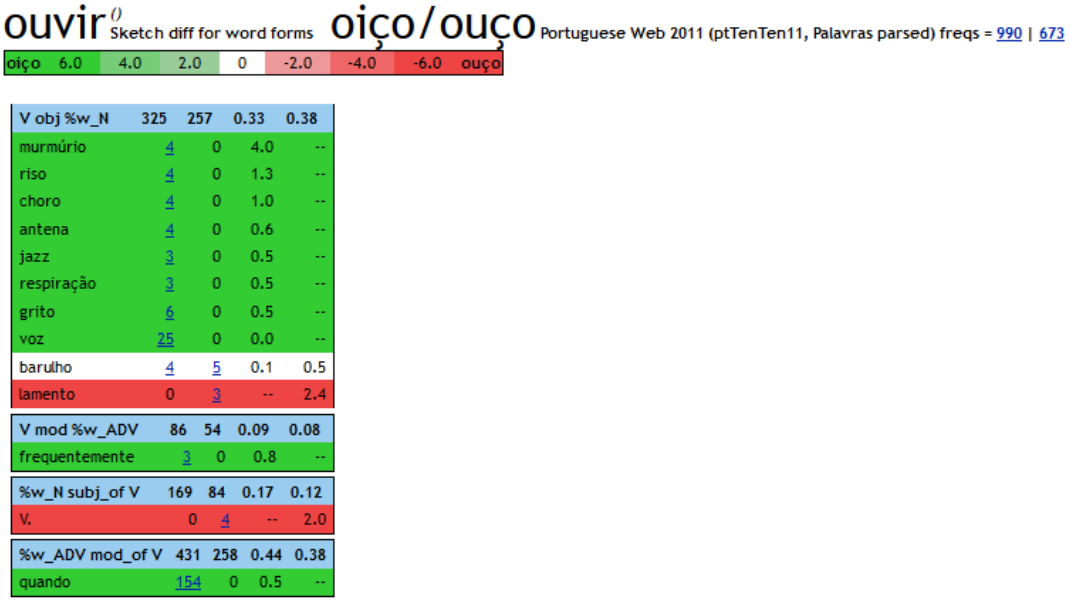


Figure 5-55 Sketch Diff search result (*oiço, ouço*)

5.2 iLex

iLex is a dictionary writing system that combines a database with a dictionary writing editor. In the case of DOPU design, automatic extraction of data from CoPEP and import into the database part of iLex was performed.

Since Chapter 9 provides a detailed description of every step of the automatic extraction procedure and Chapter 11 shows how iLex works at the database entry level, only a brief account of the program will be given here.

5.2.1 Opening a project

After opening a project, on the left-hand side of the screen there is a "Documents" panel containing a "Design" tab (Figure 5-56) with a list of the different XML schemas, and a "Look up" tab (Figure 5-57) with the list of the extracted lemmas.

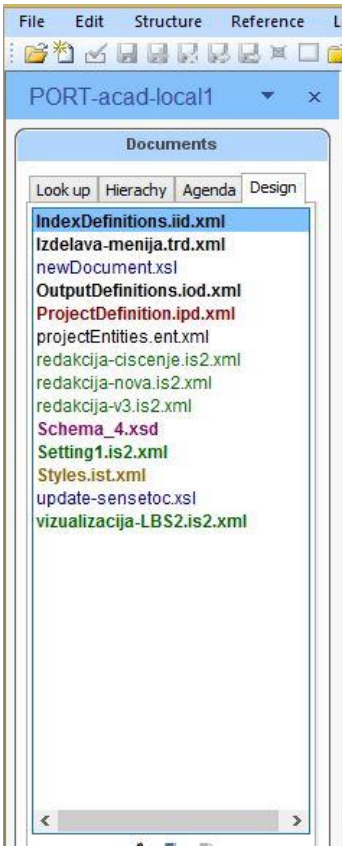


Figure 5-56 Design tab in iLex

Figure 5-57 Look up tab in iLex

Each lemma is designated by an entry-document, meaning that modifications in the structure of the entry will not apply to the other entries.

The default structure of an entry displays the following elements: headword, form, frequency, POS, sense, meaning, indicator, semantic frame, syntactic grammatical relation, collocations and examples, as shown in Figure 5-58.

Elements can be eliminated or added. For example, in Figure 5-59 a new sense was included. Attention should be drawn to indicators of the two senses, which were completed.

Alternatively, iLex offers an XML visualization of the entry, as seen in Figure 5-60. Lexicographers then have additional information to help them identify and distinguish senses, such as frequency of the collocate and language variety, besides corpus metadata, e.g., area of knowledge and year of publication.

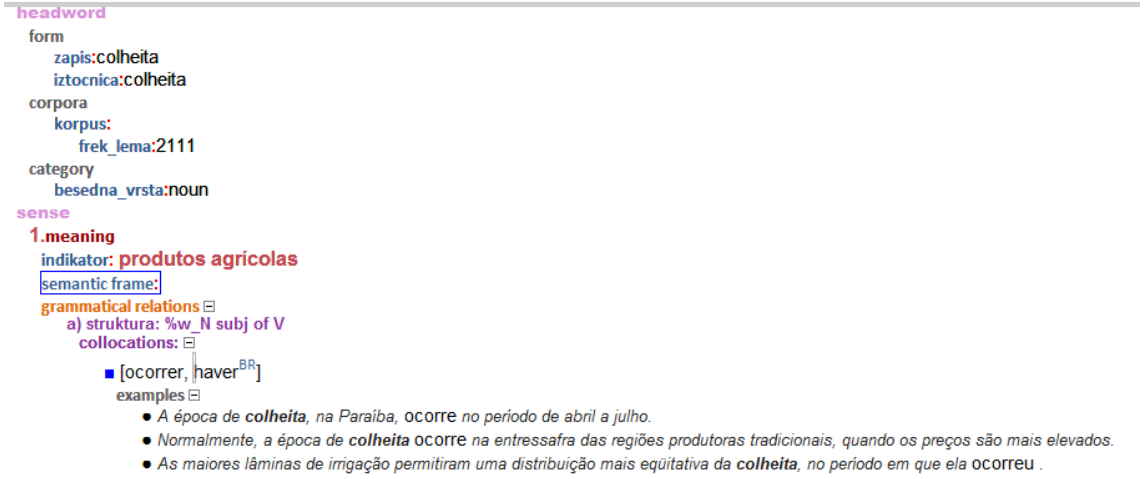


Figure 5-58 Partial view of an entry document in iLex (*ciclo*, 'cicle')

Figure 5-59 Entry with additional sense element and complete indicators

Figure 5-60 XML entry view

5.3 Summary

With the advent of new programs for corpus analysis and dictionary writing and editing, lexicographers are expected to know how to make the most of such resources. Sketch Engine is the most state-of-the-art corpus tool currently available, providing several functions and features that greatly facilitate lexicographers' routine work. Sketch Engine is fundamental for this thesis as the sophisticated computation of data it provides is automatically extracted into the dictionary writing system. Additionally, resources and tools development, as will be described in the next chapters, were all performed on the Sketch Engine. iLex is a flexible program with a user-friendly interface and large capacity for data processing. These two tools are the basis for the methodology of dictionary-making adopted in this thesis.

Chapter 6 The Corpus de Português Escrito em Periódicos (CoPEP)

DOPU's target users are students in higher education attending courses in different areas of knowledge, whose language of instruction is (Brazilian or European) Portuguese, and who thus need to read and write academic texts in Portuguese. As a corpus-driven dictionary, that means the linguistic information it portrays must be obtained from texts that reflect *the way language is used by expert writers from Brazil and Portugal in academic written production in different areas of knowledge*. Hence, the corpus required for making DOPU must have the following characteristics:

- Composed of academic written texts portraying exemplary language
- Balanced in terms of Portuguese varieties: 50% in Brazilian Portuguese, 50% in European Portuguese
- Covering different academic areas
- Synchronic
- Large in size

As shown in Chapter 3, despite the large number of existing corpora of Portuguese, none of them fully meets the lexicographical needs for making DOPU. Consequently, a new corpus, the CoPEP – *Corpus de Português Escrito em Periódicos* ('Corpus of Portuguese from Academic Journals')[159] was specially compiled for this purpose.

Interrogation of CoPEP, according to the methodology adopted in this thesis, was performed in the Sketch Engine (see Chapter 5). Development of some of the

[159] While I conceptualized and designed the CoPEP corpus, that is, while I made all the decisions regarding the composition of the corpus, bearing in mind its function for my research, thus choosing textual sources, defining parameters for domain assignment to journals, sorting out journals and issues to be extracted, and determining clean-up delimiters, my colleague José Pedro Ferreira dealt with the computational part. We worked as a team as every automatic process carried out was manually evaluated by me and served as a basis for me to indicate what was needed next. This cohesive, rigorous working system reflects my effort in making CoPEP a reliable source of data for DOPU's design, a goal that José Pedro fully embraced. For that reason, we co-authored a contribution at the Teaching and Language Corpora Conference (2016) about building CoPEP.

It is worth mentioning the reason for providing detailed information on the computational procedure conducted by José Pedro Ferreira, rather than simply indicating that a script was written for a certain phase of the compilation. Considering that this chapter might be used by other researches as a reference for corpus compilation, I understand that it would be a great contribution if programs, tools, and algorithms were explicitly exhibited.

requisites for automatic extraction, namely, Sketch Grammar (Chapter 7) and GDEX (Chapter 8), revealed some challenges in the computational treatment of CoPEP. Thus, the corpus was specially post-processed for the purpose of this research and additional adjustments in the tagger were made by the Sketch Engine team.

This chapter begins with a detailed presentation of the compilation of CoPEP, from the preparation phase (section 6.1) to the step-by-step description of the process of building it (section 6.2). It then moves on to the introduction of CoPEP and its characteristics in 6.3. Finally, section 6.4 wraps up this chapter by presenting the challenges posed by the use of CoPEP for this thesis and the measures taken to address them.

6.1 CoPEP design

Certain practical conditions were to be considered for carrying out this process: it should require no payment for journal subscriptions, texts digitization, hiring of support personnel, services, or software. And particularly important, it should be doable by a team of two people.[160]

With these conditions set, the initial step of the compilation process was the definition of texts sources that comply with the corpus characteristics mentioned above (section 4.1.1). Next, a series of operations involving texts extraction and conversion, encoding homogenization, clean-up, file name giving, areas of knowledge balancing and language variety identification was performed (section 4.1.2).

6.1.1 SciELO as the source of texts

A decision was made that the corpus would comprise texts from online journals due not only to the requirements posed by the practical conditions expressed above but also because, according to Hyland, articles "are often presented to students as good academic writing and as an ideal to be emulated as far as possible" (2008, p. 47). In other words, this kind of material perfectly meets the purpose of the corpus and the dictionary for university students I propose here.

[160] As mentioned in the previous footnote, that team was formed by this author and José Pedro Ferreira. I am also deeply indebted to Dr. Iztok Kosem, with whom I had very inspiring talks about corpus design.

Contemplation of different alternatives for journal selection has indicated that SciELO was the ideal source of texts for the corpus required for this thesis, as will be shown below.

Scientific Electronic Library Online (SciELO[161]) is a database of open access scientific journals with national collections from 15 countries, including Brazil and Portugal. It is the result of a partnership among FAPESP – the State of São Paulo Science Foundation, BIREME – the Latin America and Caribbean Center on Health Sciences Information, and other Brazilian and international institutions related to scientific communication. In 2002, CNPq – the National Council for Scientific and Technological Development (Brazil) – started supporting the project as well.

It all started with SciELO Brazil, the first national collection which became fully operational in June 1998 after a successful 14-month period of pilot testing with 10 Brazilian journals from different disciplines. Its success led to the expansion of the project, with the incorporation of new journals and participation of other countries. SciELO became:

> …a model to manage and operate electronic publications into a cooperative network of collections of increasing quality scientific periodicals with open access focused on developing and emerging countries.[162]

In addition to Brazil, other members of the SciELO network[163] with certified journal collections in regular operation are Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Spain, Mexico, Peru, Portugal, South Africa, Uruguay, and Venezuela, whereas Paraguay's journal collection is in development.

Given the key role of the source of the texts in making a corpus-driven dictionary, SciELO was considered ideal for this project due to the complete? fulfilment of all requirements mentioned at the outset of this chapter. On the one hand, SciELO hosts a great deal of current, refereed scientific journals from Brazil and Portugal covering different thematic areas, thus providing large amounts of samples of expert writing from both countries and varied subjects. On the other, it is free, online, open

[161] http://www.scielo.org

[162] http://www.scielo.org/php/level.php?lang=en&component=42&item=3

[163] As of January 2017.

access and has a common organizational structure for all national collections, facilitating the identification and extraction of textual metadata and corpus balancing.

Here I go over each one of the criteria for the creation of this corpus, explaining the extent to which SciELO meets these requirements and highlighting further strengths of this source.

a) The corpus must be composed of academic written texts portraying exemplary language

The SciELO model has very strict criteria for admission and retention of scientific journals in its national collections,[164] such as scientific content, a peer-review process, journal usage, and impact factor, which imply publication of high quality, very well-written documents in different domains. It thus seems plausible to conclude that these texts portray exemplary language.

Despite this logical conclusion, which is supported by Hyland's statement about the use of articles as models for students (see above), there have been arguments against the use of this genre for teaching academic writing under the claim that students are not required to read texts of this kind.

However, at least when it comes to the Brazilian context, statistics about SciELO have suggested otherwise. It has been shown that in July, when the Brazilian mid-year university vacation takes place, there is a drop in the number of downloads of about 10 million from the previous month, evidencing the high use of SciELO by students (Packer, Cop, Luccisano, Ramalho, & Spinak, 2014).

b) The corpus must be balanced in terms of Portuguese varieties – 50% in Brazilian Portuguese, 50% in European Portuguese – and cover different academic areas

As mentioned earlier, SciELO hosts both Brazilian and Portuguese national collections of journals, which are consequently subjected to the same strict criteria for admission and retention explained above.

Moreover, the common organizational infrastructure followed by the SciELO network has important implications for the corpus design in at least two ways. Firstly,

[164] See SciELO-Pt: http://www.scielo.mec.pt/avaliacao/avaliacao_en.htm; SciELO-Br: http://www.scielo.br/avaliacao/avaliacao_en.htm

journals in all collections follow the same classification of thematic areas, which, on the one hand, conveys an objective character to the very delicate issue of domain areas definition[165] and, on the other, facilitates the build of a balanced corpus, not only between language varieties, but also with respect to equal distribution among subjects. Secondly, an equal framework setting up the organization of the SciELO network implies the adoption of similar text markup and structure in each national website. Thus, the computational operations necessary for the identification of text metadata, and texts extraction, conversion, homogenization and clean-up should be applicable for both the Brazilian (www.scielo.br, henceforth SciELO-Br) and the Portuguese (www.scielo.mec.pt, henceforth SciELO-Pt) SciELO journal collections, hence streamlining the corpus compilation process.

c) Synchronic

Another asset is SciELO's contemporary character. On SciELO-Pt, journals publication dates range from 1997 to present, while on SciELO-Br, which also hosts complete runs of journals collections[166], dates go from 1909 to present, with the bulk of publications from 1998 onwards. Given this characteristic, SciELO has the ideal conditions to provide texts to make a synchronic corpus.

d) Large in size

In addition to all the benefits above, a large carefully designed corpus can be potentially built with texts from SciELO considering that SciELO-Br has 345 journals and SciELO-Pt, 55. Despite the reduction of the total number of usable journals to approximately 110 due to the balance-between-language-variety requirement, a rough word count estimated a corpus size reaching 45 million words. When compared to general language corpora, this is a small number. However, SciELO's compliance with the very strict design requirements set out in the building of the corpus needed for

[165] While categorization of areas of knowledge varies with indexing platforms, libraries and universities, making it very hard for corpora compilers to decide how to allocate texts under those areas, SciELO's adoption of one and only roll of subjects under which journals are integrated minimises the need of employment of subjectivity in the process of texts classification in our new corpus, thus reducing the risk of having future problems with the labelling of areas of knowledge in DOPU.

[166] According to Montanari and Packer (2014, p.71), the most important complete collection available in SciELO is "Memórias do Instituto Oswaldo Cruz", whose first issue is from 1909. It is "an international journal of biological and biomedical research published by the Instituto Oswaldo Cruz (Oswaldo Cruz Institute) […] it is today one of the most highly cited journals published in Latin America." (from http://www.scielo.br/revistas/mioc/iaboutj.htm).

making DOPU has led to the choice of using the two national collections as the only sources of texts, which constrained the corpus size, but guaranteed its quality.

In this subsection, the reasons for using SciELO as the source of the texts for our new corpus has been demonstrated. The next section describes the compilation process.

6.2 Compilation process description

The process described here took place between February and August 2016. The initial step was to gain detailed information about the journals and texts in the two national collections of SciELO so that decisions could be taken on how to continue the process of compilation according to the steps described in Biber et al., 1998; Meyer, 2002, and McEnery et al., 2006 (see Chapter 3).

6.2.1 Getting to know my sources

The first step involved identification of the journals from each national collection. Thus, automatic extraction of information (available as metadata) from the titles of each journal, unique ISSN[167] (International Standard Serial Number) identities, number of issues per journal, and language of publication allowed us to make a very rough estimation of areas of knowledge covered and subcorpora sizes.

Firstly, I manually assigned each journal to different scientific domains, following the Capes[168] classification of areas of knowledge in School of Life Sciences (henceforth CV), School of Humanities (henceforth HU) and School of Exact, Technological and Multidisciplinary Sciences (henceforth CE), as shown in Figures 6.1-

[167] The ISSN number - International Standard Serial Number – was used as a variable in the text extraction process, guaranteeing non-repetition of journals in the corpus. Furthermore, this unique identity number was included in the file names, providing ease of interoperability with the SciELO platform, and with that, prompt access to the original source.

[168] Capes stands for *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* ('Coordination for the Improvement of Higher Education Personnel') and is a foundation within the Ministry of Education in Brazil. It has created the Table of the Areas of Knowledge of Higher Education, with four hierarchical levels, from the most general– Great area – to the most specific– Speciality. In order to facilitate the activities of evaluation, which is one of its lines of action, Capes has adopted a broader categorization, with Great Areas grouped in Schools, adopting 'affinity' as the main clustering criterion. In CoPEP, this broader categorization has been adopted. For more information (in Portuguese), see: http://www.capes.gov.br.

6.3.[169] This was an attempt to gain a broad overview of the distribution of journals per area in each SciELO collection without going into more specialized subdivisions of domains.

COLÉGIO DE CIÊNCIAS EXATAS, TECNOLÓGICAS E MULTIDISCIPLINAR

CIÊNCIAS EXATAS E DA TERRA	ENGENHARIAS	MULTIDISCIPLINAR
Astronomia / Física	Engenharias I	Biotecnologia
Ciência da Computação	Engenharias II	Ciências Ambientais
Geociências	Engenharias III	Ensino
Matemática / Probabilidade e Estatística	Engenharias IV	Interdisciplinar
Química		Materiais

Figure 6-1 School of Exact, Technological and Multidisciplinary Sciences

COLÉGIO DE CIÊNCIAS DA VIDA

CIÊNCIAS AGRÁRIAS	CIÊNCIAS BIOLÓGICAS	CIÊNCIAS DA SAÚDE
Ciência de Alimentos	Biodiversidade	Educação Física
Ciências Agrárias I	Ciências Biológicas I	Enfermagem
Medicina Veterinária	Ciências Biológicas II	Farmácia
Zootecnia / Recursos Pesqueiros	Ciências Biológicas III	Medicina I
		Medicina II
		Medicina III
		Nutrição
		Odontologia
		Saúde Coletiva

Figure 6-2 School of Life Sciences

[169] The tables in Figures 6.1 to 6.3 were taken from http://www.capes.gov.br/avaliacao/sobre-as-areas-de-avaliacao

COLÉGIO DE HUMANIDADES

CIÊNCIAS HUMANAS	CIÊNCIAS SOCIAIS APLICADAS	LINGUÍSTICA, LETRAS E ARTES
Antropologia / Arqueologia	Administração, Ciências Contábeis e Turismo	Artes / Música
Ciência Política e Relações Internacionais	Arquitetura e Urbanismo	Letras / Linguística
Educação	Ciências Sociais Aplicadas	
Filosofia / Teologia	Direito	
Geografia	Economia	
História	Planejamento Urbano e Regional / Demografia	
Psicologia	Serviço Social	
Sociologia		

Figure 6-3 School of Humanities

As expected, given the great difference in time span coverage of each national collection (see 4.1.1 above), SciELO-Pt is much smaller than SciELO-Br. As of February 2016, there were 55 journals in SciELO-Pt. The number of issues ranged from 1 to 75 per journal, totalling 965 issues with the following distribution in schools: 457 (HU), 426 (CV), and 82 (CE), whereas SciELO-Br contained 345 journals on the website. The number of issues per journal varied from two to 440 issues, amounting to 18,270 issues in total. Here, the School of Life Sciences is predominant: 11,454 issues, followed by 9,054 issues from Humanities and 2,209 from the School of Exact, Technological and Multidisciplinary Sciences.

Before moving on to text extraction, a decision was made that the language variety in each publication was by default assumed to be that of the country of publication, that is, texts extracted from SciELO-Pt made up the European Portuguese subcorpus (PT), whereas those extracted from SciELO-Br comprised the Brazilian Portuguese subcorpus[170] (BR). Consequently, the final configuration of the corpus in

[170] That is a key point because, although the assumption is made that contributions in each one of the national collections should have been written in the language variety of the respective country of publication, tradition allied with globalization have proved otherwise. On the one hand, not only is it quite common for Portuguese academics to be published in Brazil and vice-versa, but also for contributions resulting from team-work composed of Brazilian and Portuguese researchers to be published in either country. On the other, our globalised world has given place to an ever-growing internationalization of universities and research centres, with both students and researchers publishing in their country of origin but also of residence. That means national collections cannot be taken for granted as representative of language variety. Refinement of variety balance was done at the end of the compilation process and will be shown later.

terms of size and balance was determined by the smallest collection, which is the SciELO-Pt. Language variety identification, which was intended to enhance accuracy of subcorpora balance, was performed at the very end of the process, as will be shown later.

6.2.2 Building the corpus

The first part of this section describes the process of corpus building based on XML files extracted from SciELO national collections. Although extraction of web pages is in XML format, due to serious problems of XML inconsistency in SciELO, the only alternative was to extract HTML files. Thus, the second part displays this process that, on a positive note, incorporated enhancements learnt from the first extraction. Due to this unpredicted issue with the website structure, the process of corpus compilation took longer than expected.

6.2.2.1 XML extraction

With time span coverage of the corpus set from 2000 to 2015, schools of knowledge (HU, CV and CE) assigned to journals, SciELO-Pt defined as the first collection to be extracted, and the acknowledgment that the URLs for articles in SciELO contain a unique identifier based on the pattern "ISSN-ISSNYEARISUENUMBR", computational procedures[171] began. These were the steps taken:

a) Taking as variables the above-mentioned pattern, the defined time span, the assigned schools of knowledge and the definition that XML files written in Portuguese were to be extracted, links were over-generated using regular expressions.
b) These links were downloaded by using a weget app in a for loop bash based on this kind of command:

    wget -nv -nc -O <pasta_destino/ficheiro_destino.xml> <URL>

c) Links were placed under SciELO-Br or SciELO-Pt folders and under HU, CV or CE folders referring to schools of knowledge.

[171] As mentioned earlier, computational procedures were performed by my colleague José Pedro Ferreira.

d) Raw XML articles from SciELO-Pt were extracted using XSLT (xsltproc engine under Debian) and W3C's main translational XSL.
e) Each article was automatically saved as a file whose name contained the ISSN, year, issue number, article number and school of knowledge.

The extraction script running on SciELO-Pt had several problems, stopping many times and making extraction take a very long time. Meanwhile, I selected journals and number of issues per journal to be extracted from SciELO-Br, aiming at keeping balance regarding school and topic equivalence with SciELO-Pt.

The vast majority of the journals presented direct correlation, for instance, *Psicologia* (SciELO-Pt) and *Fractal: Revista de Psicologia* (SciELO-Br). However, others did not. In that case, I manually analysed the description of the journal in SciELO-Pt and tried to find an equivalent in SciELO-Br. Table 6.1 shows two examples as an illustration of the process of correspondence search, pointing out the decision taken and the justification.

Table 6.1 Illustration of the process of correspondence search

School of Knowledge	Scielo.pt	Scielo.br	Justification
CE	*Revista de Gestão Costeira Integrada* (16 issues)	*1. Revista Brasileira de Oceanografia* (5 issues) *2. Revista Ambiente & Água* (11 issues)	No direct correspondence of the topic. Insufficient number of issues in journal with nearest topic (1). Selection of a second journal (2) with a broader, generic topic.
HU	*Revista Diacrítica* (12 issues)	1. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* (4 issues) 2. *Estudos de Literatura Brasileira Contemporânea* (4 issues) 3. *Kriterion: Revista de Filosofia* (4 issues)	Pt journal covers three topics: linguistics, literature and philosophy. Three publications per year: one issue, one discipline.

As can be seen, notwithstanding the same school of knowledge classification, efforts were made to obtain a maximum correspondence as possible in terms of area/topic. However, there were some cases in which this kind of equivalence was not found. This happened, for instance, with a journal of computational sciences in the Portuguese collection, containing 2,648 files (texts). Given that SciELO-Pt is already a

small collection of texts in the first place, I did not want to eliminate any journals thereby losing further data. I then decided to keep school of knowledge correspondence only, even though the areas would not be similar. A journal on production engineering was then selected from SciELO-Br.

When extraction of SciELO-Pt articles was performed, we realised that XML metadata contained information on great areas of knowledge for each journal, following Capes' classification presented in 4.2.1 above. Thus, this information was also extracted from both national collections. Table 6.2 summarises the great areas used in our corpus.

Table 6.2 Schools and Great Areas of knowledge in CoPEP

Schools	***School of Humanities (HU)***		***School of Life Sciences (CV)***		***School of Exact, Technological and Multidisciplinary Sciences (CE)***	
Great Areas	Human Sciences (Hu)	Applied Social Sciences (Ap)	Health Sciences (He)	Agricultural Sciences (Ag)	Engineering (En)	Earth and Exact Sciences (Ex)

As some journals were classified under more than one great area (e.g., *Revista Portuguesa de Ciências do Desporto* (Health Sciences, Human Sciences)), an attempt was made to provide objective criteria for determining only one discipline per journal. I then followed these steps:

i. As the *Fundação para a Ciência a a Tecnologia* (FCT) in Portugal broadly corresponds to Capes in Brazil, I verified to which great area journals were assigned in FCT and adopted this area. For instance, the *Revista Portuguesa de Ciências do Desporto* (Health Sciences, Human Sciences) belongs to Health Sciences in FCT, thus, in my corpus, this journal is classified as great area: He; school: CV.

ii. If journals were not displayed in FCT, ISI Web of Knowledge classification was checked. For instance, the journal *Nascer e Crescer* is classified in SciELO-Pt as Applied Social Sciences, Biological Sciences, Health Sciences, Human Sciences. According to ISI Web of Knowledge, it belongs to Paediatrics, which is a specialization of Health Sciences.

iii. If neither FCT nor ISI provides a clear classification, then the school of the corresponding journal in the other collection is adopted. For instance, the

Cadernos de Estudos Africanos (Applied Social Sciences, Human Sciences, Linguistics, Arts and Humanities) from SciELO-Pt were attributed to two corresponding journals in SciELO-Br, *Afro-Ásia* and *Estudos Afro-Asiáticos*, which belong to Human Sciences. Thus, *Cadernos de Estudos Africanos* was assigned Human Sciences great area.

Another unexpected result from SciELO-Pt articles extraction is that some journals, despite originally marked as English-only due to titles in English confirmed by random read, did have a few articles in Portuguese. These were added, and new corresponding articles had to be found in SciELO-Br, for which I followed the same manual analysis of journals described above. In the end, only one journal was exclusively written in English and was thus discarded.

Finally, additional results from extraction showed that in many cases where the language was determined to be Portuguese, texts were only available in PDF format (e.g. http://www.scielo.mec.pt/scieloOrg/php/articleXML.php?pid=S0872-19042003000100001&lang=pt). This was discovered as the line Texto completo disponível apenas em PDF. was identifiable in the XML. I then opted for the exclusion of these documents from the folder with extracted texts, however, keeping a record of the file names (1650 files in total) in a different folder. This choice was made as conversion of PDF files to txt format requires manual review of individual files due to recurrent problematic results (e.g. two-column-per-page articles completely lose their internal structure).

After this series of manual actions, journals and number of issues to be extracted from SciELO-Br were performed, and the extraction script began to run. It should be mentioned that, as with SciELO-Pt, there were many server down events and intermittent interruptions.

Once the Br and Pt raw subcorpora were complete, additional computational procedures were performed:

f) Automatic clean-up of extracted XML files: author bibliographical information (name, affiliation, email, position), abstracts, keywords; images; tables; figures; charts; references; extra information (copyright information, publisher's address, acknowledgements; received on, accepted on).

g) Substitution of school folders with school and great area in name files. Thus, every file was batch renamed based on fixed-position codes, retaining the unique identifier used in Scielo's URLs for future reference. For example, in BRCVHe0104-42302008000100023:

BR: country top-level domain for the article's variety of Portuguese

CV: school of knowledge

He: great area

0104-4230: ISSN

2008: publication year

0001: issue

00023: article number within the issue

h) Encoding homogenization was performed. Files in Latin1 / ISO-8859-1 encoding were converted to UTF-8.

When checking on the transformations resulting from the clean-up, it was noticed that many files were not valid XMLs, suggesting inconsistency of SciELO organizational structure. As alternative solutions were being attempted, I was provided with the cleaned and converted files (46,935 in total) in order to analyse a sample of texts and evaluate the extent of the problems. It should be mentioned, though, that this analysis was not expected to yield statistics, but only to verify whether the issues were spread over all files.

The process of manual analysis was performed as described below.

i. Description of issues under inspection (identified during transformation results check-up):
   - Inadequate cleaning: presence of XML markers
   - Empty files: no textual content
   - Useless files: no qualified textual information, e.g. abstracts in Portuguese and other languages (Spanish or English), biographical data, references, keywords, etc.

ii. Data organization:

1. Files were grouped according to language variety, school and great area, and the total number of files for each group was recorded:

Table 6.3 Number of files per language variety subcorpora and great areas of knowledge

Br corpus	Pt corpus
CVAg (4.115 files)	CVAg (758 files)
CVHe (14.966 files)	CVHe (3.321)
CEEn (1.345 files)	CEEn (178 files)
CEEx (4.083 files)	CEEx (189 files)
HUAp (2.959 files)	HUAp (1.131 files)
HUHu (9.733 files)	HUHu (3.788 files)

2. In File Explorer, files of each group were ordered by size, with the largest file at the top. One file was picked approximately every 10KB size-difference, until the minimum size of 10KB. The determination of this value resulted from manual analysis of smaller-sized files, showing only junk, while 10KB files sometimes corresponded to editorials.

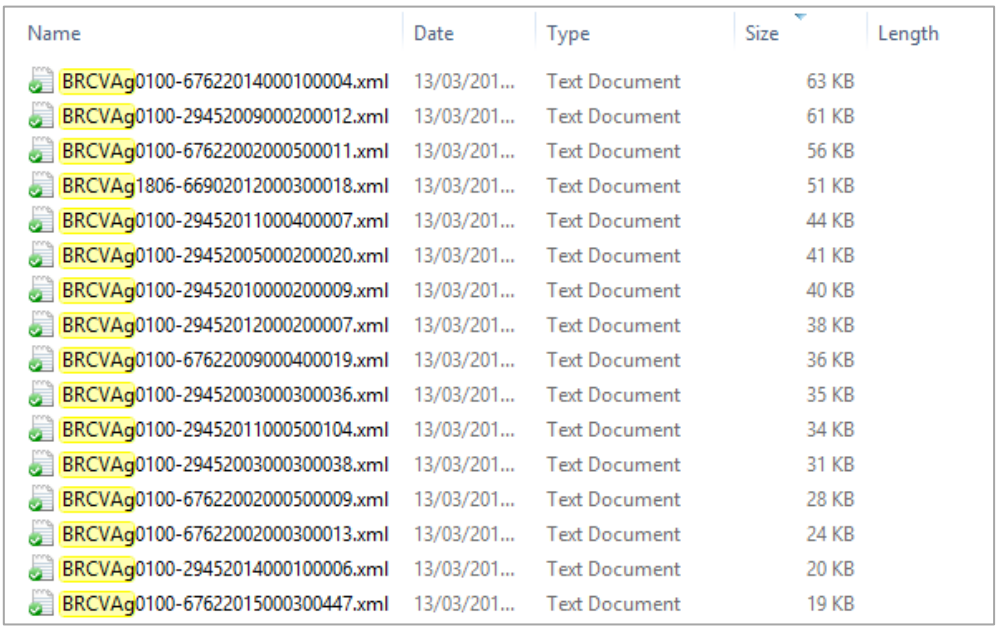


Figure 6-4 Files for selection

3. Sampled files were grouped in folders. Each folder was uploaded into CountAnything[172] program for words and characters counts. Results were saved as text files;

4. Results were imported into an Excel spreadsheet, ordered by size.

Table 6.4 Files in Excel spreadsheet

	Size	Words	Chars
BRCVAg (4.115 files)			
BRCVAg\BRCVAg1806-66902016	10KB	1.190	9.738
BRCVAg\BRCVAg0100-67622015	15KB	1.667	14.205
BRCVAg\BRCVAg0100-29452014	20KB	2.525	17.681
BRCVAg\BRCVAg0100-29452003	31KB	3.836	26.676

iii. Qualitative analysis:

1. I opened one file at a time and examined:

- coding conversion

- presence of XML markers

- presence of body of text (none, incomplete, complete,);

1.1 As to body of text, the notes taken were:

- if no body of text was present: no use

- if complete body of text was present: full article

- if incomplete body of text was present: incomplete

1.2 If journal section was identifiable, the name was recorded between parentheses.

iv. Findings:

The presence of encoding errors in almost all files, despite the previous conversion from original ISO-8859-1 encoding into UTF-8 and XML markers in all files. In addition, complete texts were not clean, with the presence of title, abstract(s), keywords, information about the author, references, among others. Figure 6-5 illustrates the content of texts that were analysed:

[172] http://ginstrom.com/CountAnything/

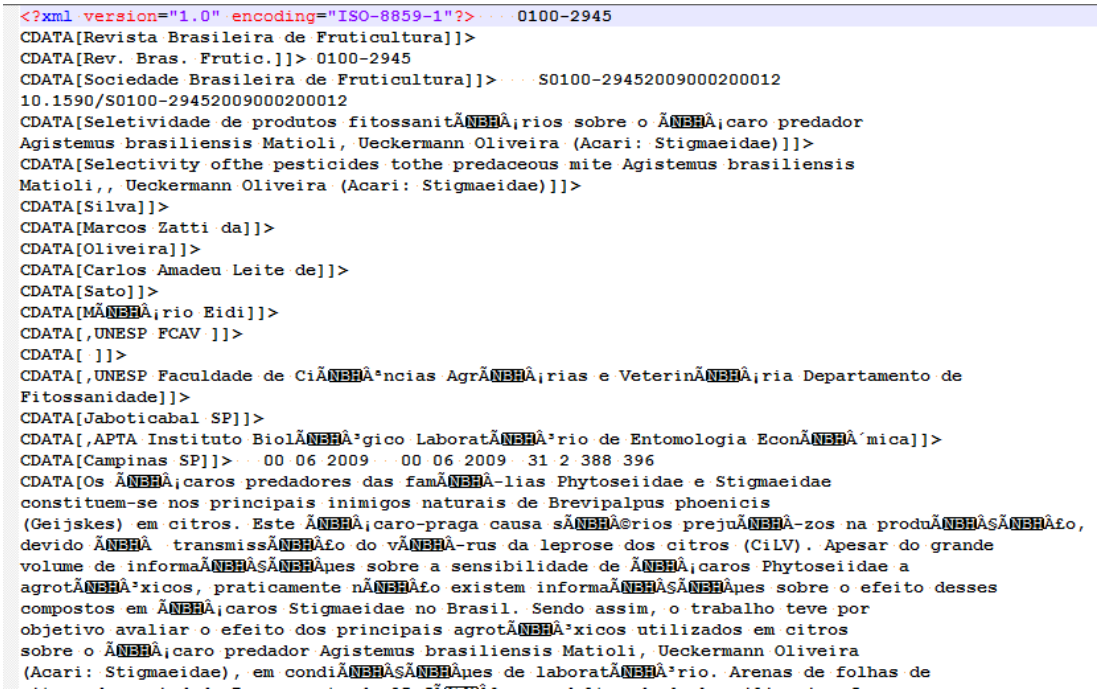


Figure 6-5 Example of the result of an XML extraction after clean-up

The total number of files in this first provisional corpus was 46,935 files. There were 44,544 files with sizes between 0 and 10 KB and 2,011 files sized between 11KB and 21KB. In the examination, 23 files out of 27 "no-use" files were smaller than 21KB, suggesting that only about 380 articles may contain texts.

v. Conclusions:

This analysis showed that poor cleaning and lack of textual content in files are pervasive in the whole corpus, indicating that XML extraction was not viable and a new solution must be found. It was thus decided that HTML extraction should be performed instead.

6.2.2.2 HTML extraction

After contact with the IT person responsible for SciELO, it was confirmed that organizational structure was not consistent, with problematic XML files.[173] The decision was then to extract HTML files.

[173] Santos and Packer's (2014, p.83) explained that, until 2013 "files received from the journal publishers were converted to plain text coded in HTML format (HyperText Markup Language) to be marked-up according to the SciELO SGML structure, and then stored in a database for online publishing and distribution". In 2013, the XML format replaced the SGML structure. This suggests that this transition from one format to the other might have been the reason for such problematic organizational structure. Nevertheless, it can be said that this issue should be overcome once the process of transformation is finished.

Despite this unexpected setback, HTML extraction benefited from lessons learnt in the first process, which enhanced and helped to accelerate the procedure of corpus building. It should be stressed that, notwithstanding extraction of valid files this time, the SciELO server was still problematic. Hence, various server down events led to inconvenient interruptions of script runs, causing more unfortunate delays.

The computational procedure was performed on SciELO-Pt as described below:

a) Files were obtained using a PHP script, automatically naming, pre-cleaning and normalising them using a DOM parser;

The same parameters were used: years 2000 to 2015, ISSN, issue number, article number, language source, school and great area.

b) Some files were put aside (the ones clearly indicating PDF-only availability) using a grep-based bash script;

In this step, 1650 files whose texts were only available in PDF were excluded.

c) Regex-based sed commands were used for cropping and trimming the files; pointers were variations on: upper limit – "keywords"; lower limit – "references";

d) Final HTML cleanup was done using an asciinator (html2text);

e) Finally, remaining character entities were recoded using GNU recode.

As in the previous extraction, I was provided with extracted files, this time the SciELO-Pt subcorpus (7,740 files), in order to analyse a sample of text and evaluate the results of this new procedure. Meanwhile, the script was still running on SciELO-Br.

The analysis was performed this way:

i. Files with 1KB, 2KB and 3KB sizes were all opened. 1KB-sized files were junk. Files with 2KB sometimes displayed short texts that should be maintained, for instance, "*carta ao Director*", "*nota introdutória*", "*posters*", "*notícias*", "*editorial*", and sometimes useless content. These last kinds of files usually contained a correspondence address, title, author's name, affiliation, abstract, email, among other irrelevant information for this corpus purposes, together with the sentence "Full text only available in PDF format." Some files with 3KB

consisted of relevant information, while others were either incomplete texts or simply abstracts in other languages with other information.

ii. In File Explorer, files were ordered by name. Then, I selected five files, went down the list some 200 files and selected another five files, repeating this process until reaching the end of the list and avoiding files smaller than 3KB. In the end, the sample contained 195 files.

iii. I opened the files one by one and took note of the different kinds of unwanted information found or cases of wrong text trimming (missing beginning or end of the text). The coding I used can be seen in Table 6.5.

Table 6.5 Codes created for HTML extraction results qualitative analysis

AB = ABSTRACT
Ad= address
Af= affiliation
Ag= AGRADECIMENTOS/Agradecimentos (acknowledgments)
AR= article/book reference (in cases of reviews)
Au= author's name
bar==========
CC= [Creative_Commons_License]
Ec=encoding issues
Ed= Editorial
Em= email address
Ib= incomplete text - missing the beginning
Ic = incomplete text -interrupted conclusions
img= [/img/revistas/cpm/v29n1/29n1a02t1.jpg]
KW= keywords
PC= Palavras-chave (keywords)
PDF = Full text only available in PDF format.
RC= Responsabilidades éticas/confidencialidade dos dados/conflito de interesses (ethical responsibility/data confidentiality/interest conflict)
Re= RESUMO
RF= Referência bibliográficas (references)
RP= Recebido para publicação em (received for publishing) /data de rece(p)ção (date of reception)
Sp= beginning of the text in Spanish. From Resumen to Descriptores (inclusive)
SU= Summary
Ta= scrambled table
Ti= title

iv. Findings

The HTML extraction was very successful as all files were plain texts, with no markups of any kind. Clean-up was 100% effective in 8 out of 195 files, with different levels of efficiency in the remaining ones, varying from some common, expected kinds of unwanted information to those that can be eliminated through additional automatic clean-up (differently from the XML extraction, where texts were simply junk).

Encoding conversion to UTF-8 was correct, with only failed transformation in the case of a few symbols, as seen in Table 6.6:

Table 6.6 Failed symbols conversion

&amp	&
&ndash	- (dash)
&rsquo	' (apostrophe)
&ldquo	word/sentence
&rdquo	" "

Another significant result of the analysis is that I found a pattern that could be used in the new automatic clean-up. It refers to the presence of "[Creative_Commons_License]" at the end of the text. Sometimes it is followed by unnecessary information, namely, institution, institution address, a place-holder for image and an email address. For instance:

> [Creative_Commons_License] Todo o conteúdo deste periódico, exceto onde está identificado, está licenciado sob uma Licença_Creative_Commons
> Instituto de Ciências Sociais da Universidade de Lisboa
> Av. Professor Aníbal de Bettencourt, 9
> 1600-189 Lisboa
> [/img/pt/e-mailt.gif]
> clara.cabral@ics.ul.pt

Whether followed by some extra strings of characters or not, the "Creative Commons" line was found to be an effective lower cut-off point.

v. Conclusions

It was decided that 1KB-sized files could be safely eliminated as they were junk. Moreover, it became apparent that some extra cleaning should be performed in which observations on sizes and patterns (PDF, Creative Commons License) could be incorporated as parameters.

6.2.2.2.1 Second HTML extraction

The results of the analysis were sent to the computational expert, who implemented them in a new automatic clean-up script. The new exclusion parameters were:

- files smaller than 2.5KB
- Creative Commons pattern
- files smaller than 5KB containing “in? PDF?<" or "em? PDF?<"
- a larger time span: 1997 to present

The decision to extend the time span to the year of the oldest publication in SciELO-Pt resulted from the significant size reduction of this subcorpus due to the considerable number of “PDF-only” files.

A second clean-up was performed on SciELO-Pt and SciELO-Br, which had been fully extracted by this time. Table 6.7 shows total numbers for the treated data, that is, after the parameters above had been applied. It should be noted that another 706 files with an indication of PDF-only availability were excluded from SciELO-Pt, totalling 2,356 files from this subcorpus that, in principle, cannot be used (as mentioned in 6.2.2.1 above).

Table 6.7 Number of tokens per language variety subcorpus and great area of knowledge

	Pt tokens	**Br tokens**
HUHu	12 408 219	44 433 173
HUAp	2 959 848	11 686 493
HU	15 368 067	56 870 606
CVHe	6 163 331	20 744 523
CVAg	1 196 919	11 330 446
CV	7 360 250	32 134 231
CEEn	332 449	7 549 102
CEEx	602 955	14 509 119
CE	935 404	23 821 008
Total	23 663 721 6632 files	114 334 616 25 619 files

As can be seen, the SciELO-Br subcorpus is much larger than its counterpart, requiring a careful process of balancing. Based on the number of tokens per great area, files (in order of size) were selected first until reaching an approximate total count of

that in SciELO-Pt. This first version of the corpus[174] had 47 341 296 tokens: PT: 23 663 721 and BR: 23 677 575. Nevertheless, it should be remembered that these subcorpora refer to the source of publication, not language variety.

While the corpus was being compiled, I was also devising a sketch grammar for academic Portuguese (see Chapter 7). Thus, as I found errors resulting from poor clean-up in concordances (e.g. the presence of a whole paragraph in another language, keywords, abstracts, acknowledgments, etc.) in the Sketch Engine, I kept a record of the filenames and searched for them in the corpus. I then manually reviewed all of the texts, correcting other errors that occasionally came up, namely, second titles in English, author names, "glued" words, or simply discard files with incomplete texts and additional "PDF-only" files.

In addition to the manual clean-up, I also took note of files written in a language variety contrary to that of the publication for future filenames modification, for instance:

> #articles in BrPt found in Scielo.pt
> PTCEEx1646-88722014000100005.html.txt
> PTCVAg0871-018X2014000300005.html.txt
> #articles in EuPt found in Scielo.br
> BRCVHe0034-71672014000300360.html.txt
> BRCVHe0034-71672014000600913.html.txt
> BRHUHu0002-05912014000100002.html.txt

The procedure for language identification should have been performed on this cleaner version of the corpus. However, as the total number of tokens in SciELO was reduced even more due to the results of manual evaluation of the texts, I decided that the conversion of PDFs should be attempted since 2,356 files could potentially provide a great deal more data.

6.2.2.2.2 PDF conversion

Ideally, PDFs would have been automatically converted, leaving me to manually analyse the results. However, due to some delays in the computational procedure, I started the conversion process manually. It entailed downloading the PDFs one by one,

[174] I named this version "the Pre-final corpus". It was used for the first GDEX configuration development, the devising of the Sketch Grammar, and the experiment of automatic data extraction from the corpus and import into DWS.

then uploading them to a free online PDF to txt converter and saving the results, that is, saving the files in txt format.

Having a list of the links for the online PDF versions of 1,650 articles from SciELO-Pt that had been discarded with the first HTML extraction, I opened each one in order to verify conversion suitability. Given that preliminary conversion tests with two-columns-per-page papers showed that this kind of structure was lost in the txt version, I created a workflow criterion in which only one-columned papers were candidates for conversion, thus avoiding unnecessary work.

I then followed the procedure mentioned above, resulting in 69 txt files, which were also manually cleaned, and included the elimination of titles in other languages, abstracts (in Portuguese and other languages), author information, keywords, acknowledgments, references, tables, images, submission and publication dates, Creative Commons Licenses, and other occasional kinds of irrelevant information. Additionally, other PDF-typical elements, namely, journal layout markers like headers and footers with author names or journal titles, page numbers, and footnotes were also manually eliminated.

Even though these texts were published in SciELO-Pt, a manual analysis showed that 18 were in fact written in Brazilian Portuguese. These filenames were recorded and a list of the links of 733 remaining PDFs for extraction was created. These PDFs were the total number left after manual conversion of 69 texts and elimination of 848 double-columned papers.

At the same time, sketch grammar was being improved, so I had the opportunity to perform further manual cleaning in the corpus (in the same manner as described earlier), discarding additional texts written in other languages and eliminating double titles, references and the indication of PDF-only texts from the texts. This new modified version of the corpus was saved to be treated later.

I was next provided with 638 automatically converted txt files for manual cleaning. Initially, I put many of them aside due to irregular text structure, focusing on convertable files pertaining to great areas with fewer texts, such as Agricultural Sciences, Engineering, Exact Sciences and Life Sciences. However, among those previously avoided puzzle-like texts, there was a considerable amount from Exact

Sciences and Life Sciences. Hence, despite the hard work, I chose to include them in the manual review and reorganise their internal structure taking as a reference the original PDF version.

In the end, I manually analysed and cleaned 213 files, 34 of them written in Brazilian Portuguese and 179 in European Portuguese.

Having a new, cleaner version of the corpus, and an additional 213 files, it was time to move on to the final phase: the creation of language variety subcorpora, new file naming and balancing.

6.2.2.2.3 Corpus building final phase

The final phase consisted of creating a Brazilian Portuguese subcorpus and a European Portuguese subcorpus, renaming the files with the addition of a code for language variety, and balancing the corpus according to the total amount of words per subcorpora and per great area (i.e. school).

6.2.2.2.3.1 Sorting out the subcorpora texts

Until now, the corpus was divided into two subcorpora according to the publication sources of the texts, that is, SciELO-Br and SciELO-Pt. As explained earlier, a direct correspondence between the place of publication and written language variety cannot be made in the case of the Portuguese language. Thus, it was planned to have these currently source-based subcorpora analysed in order to confirm the variety in which texts were written. Files with a divergence between source and variety were moved to the other subcorpus. In the end, two language-variety subcorpora were sorted out.

For the analysis of each subcorpora, a decision was made to adopt a procedure for identification of language variety of the texts. This was carried out in two phases for each subcorpus: 1) automatic identification of BP and EP variants in texts written in EP and BP, respectively; 2) manual confirmation of a correspondence between the language identified and language of writing.

It should be noted that automatic identification of language variety is not a foolproof method due to a number of reasons. Firstly, there are cases when texts co-authored by multinationals display two varieties simultaneously. Secondly, some authors apply orthography inconsistently, randomly varying between BP and EP.

Finally, sometimes supposedly deviant spellings are, in fact, accepted forms in other varieties. Given that, the results of automatic language variety identification should be considered only indicative of the dialect of writing, not a definitive conclusion. This is why the procedure followed in this thesis involved a second phase, in which I manually reviewed the texts to confirm the automatic indication of variety.

At this point, it is important to bear in mind that the 213 manually analysed PDF-converted files from the previous section had not been added to the subcorpora yet. As they had been sorted into language varieties during manual evaluation of conversion results, it was not necessary to include them in the procedure described below. They were added to the corpus only after the new subcorpora were defined.

Beginning with the SciELO-Br subcorpus, these were the steps that were followed:

a) Running a stoplist with EP variants

A stoplist comprising EP variants (e.g. *excepção*, *facto*) was used. Filenames of texts containing at least one occurrence of any of the EP variants were saved in a list. In the end, the candidate EP texts list comprised 82 texts, varying from 40 to 1 occurrences per text.

b) Running an EP phonological endings stoplist

Next, a script with EP phonological endings (e.g., *émico*, *ómico*) was run. As above, filenames containing such endings were saved, totalling 101 files, varying from 44 to 1 occurrences of EP phonological endings.

c) Merging the two stoplists

The result was a candidate list containing 149 filenames.

d) Manual review

The purpose of the manual review was to confirm a correspondence between language variety identification and the real language of writing.

I sorted the 149 filenames in order of frequency and read each text until I reached rank 50, corresponding to the last text with three occurrences of EP variants. I then reviewed an additional 16 texts with two occurrences and concluded they were all

false positives, that is, the texts were indeed written in Brazilian Portuguese. Thus, files in the remaining tail were determined to be BP.

The manual review involved a mixture of my being a speaker of Brazilian Portuguese as a mother tongue who is familiar with distinctive features of European Portuguese, specifically in the context of written academic Portuguese, together with the identification of elements characteristic of a certain variety or belonging to one culture or another, such as (among many others):

i) For confirmation of texts written in European Portuguese:
   - *facto*;
   - *registo*;
   - acute accent with verbs in the simple past (*habituámo-nos*);
   - *acção* before 2009;
   - *ideia* before 2009;
   - collocation *deitar abaixo*

ii) For confirmation of texts written in Brazilian Portuguese:
   - *aspecto* and *ruptura* after 2009
   - *européia*, *idéia* (accented diphthong)
   - *de fato*
   - *tese de doutorado*
   - *bolsista Capes*
   - *diaeresis (ü)*

In the end, out of 149 EP candidate texts, only 20 were written in European Portuguese.

One interesting finding resulting from this review was the identification of "hybrid" writings after 2009, that is, with orthography simultaneously following AO90 and AO45. Another revealing observation refers to the presence of "misleading" paper titles, which had been transformed into BP, while the article was in fact written in EP (e.g., the title: *Índices plaquetários em indivíduos com doença hepática alcoólica crônica*).

The steps were followed in the same manner with the SciELO-Pt subcorpus, the only difference being the performance of a slightly modified method for manual inspection of the results, as will be explained below.

It is notable that the automatic identification of BP variants in the SciELO-Pt subcorpus yielded a list with 921 filenames. In comparison with the list of 149 texts above, such a huge difference suggests more Brazilians publishing in Portugal than the other way around.

Given the unmanageable size of the list, it was decided to set five as the cut-off point, meaning that all files with six occurrences or more of BP variants were determined to be BP (328 files in total) while five occurrences or less of BP variants were considered EP (593 files). However, as five is quite a high threshold, it was decided that further manual analysis should be undertaken, although not at this point due to time constraints.

Lastly, it should be noted that, while the processes above were being performed, I found sparse occurrences of cases like this: “ocidentalização da Amazônia” A cleaning script for conversion into missing entities was soon devised (based on W3C and WebStandards' tables), fixing the problems straightforwardly.

6.2.2.2.3.2 File renaming

Since the results from the section above confirmed that source could not be considered equivalent to language variety, files had to be renamed in order to accommodate this new piece of information.

The solution was to include two extra letters at the beginning of the name, corresponding to the language variety, leaving the following two to represent the source of publication. In the end, there were four possible combinations (bold and underlining are used here only to elicit the motivation of the codes):

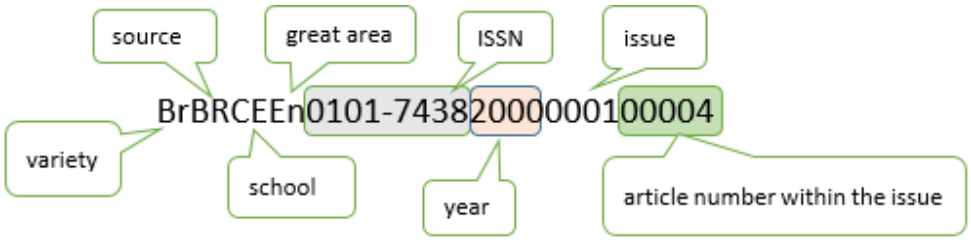
BrBR: **Br**azilian Portuguese published in SciELO-Br

EuBR: **Eu**ropean Portuguese published in SciELO-Br

EuPT: **Eu**ropean Portuguese published in SciELO-Pt

BrPT: **Br**azilian Portuguese published in SciELO-Pt

Filenames in the corpus contain all the metadata necessary for an advanced corpus search:



It should be mentioned that, at this point, the PDF-converted (213) files were added to the corpus and sorted into language variety subcorpora according to manual identification that had been previously done.

6.2.2.2.3.3 Balancing

a) Corpus first version

As known, the EP subcorpus was much smaller than the BP subcorpus. For balancing, the following exclusion criteria were followed:

- texts older than the year 2000
- texts with mismatching variety and source
- smaller files

Finally, after sorting out the subcorpora texts, renaming the files, and proceeding with the process of balancing, the first version of the *Corpus de Português Escrito em Periódicos* – CoPEP ('Corpus of Portuguese from Academic Journals') was ready:

Table 6.8 number of texts and words in CoPEP

	Whole corpus	Brazilian Portuguese	European Portuguese
Texts	10.491	4.780	5.711
Words	44.072.203	22.036.438	22.035.765

b) CoPEP

The decision to further improve the corpus led to manual analysis of the BrPT files in CoPEP_v1. Four files out of 113 were found to be actually written in EP, so they were renamed. I took advantage and carried out some more manual cleaning, removing double paper titles, abstracts, keywords, author data, among other unwanted information.

Furthermore, some more analysis was carried out on the files that had been automatically determined to be EP due to the occurrence of five or less BP variants in the SciELO-Pt subcorpus. The purpose was to manually confirm a match between language identification and language of writing. The first analysis round resulted in the identification of 51 files that were in fact written in BP, which were then renamed. In other to accelerate this process, I suggested having the remaining files undergo a reversed language identification procedure. That is, this time, EP texts were inspected for evidence confirming source and variety match. Thus, EP variants and phonological endings were run against these files.

After balancing was performed once more, a script was run for extracting filename information and converting it into metadata headers in the texts in order to facilitate text type differentiation for corpus searches in the Sketch Engine (or any other corpus tool management with this function), as shown below.

```
<doc variety="Br" source="BR" school="Ex-Tech-Multi Sciences" great area="Engineering"
issn="0101-7438" year="2000" issue="0001" article_num="00004">
Algoritmo de programação de máquinas individuais com penalidades distintas de
adiantamento e atraso


1. INTRODUÇÃO
Desde o início da difusão de princípios do JIT (Just-In-Time) ' que pode ser
considerado como um sistema de administração industrial relativo ao estoque ' ,
tem crescido a importância da diminuição do estoque no processamento de
produtos. Estamos fazendo mais um esforço nesse sentido, considerando como
característica-chave de nosso trabalho um dos elementos mais importantes do JIT
```

Figure 6-6 Part of a text with header from CoPEP

As is widely known, corpora can be continuously improved. Hence, a limit must be set. The painstaking procedure of corpus compilation described here has led to a working corpus that fully fulfils the purpose of this thesis. In consequence, further perfecting will not be pursued for now.

It should be noted that an additional step might be taken in the process of corpus compilation before it is uploaded into a corpus query tool: corpus annotation (e.g. McEnery et al. 2006, pp. 71-76). For the present project, where Sketch Engine played a central role, CoPEP annotation was carried out in the corpus tool with the default tagger for Portuguese corpora, which is Freeling v3[175](see Chapter 5). It is relevant to point out

[175] At the time of writing.

that Freeling is a free, fully corpus-query-system-independent open source tool, meaning that it can be employed regardless of any other resource.

The next section presents CoPEP in more detail.

6.3 The Corpus de Português Escrito em Periódicos -CoPEP

CoPEP contains 9,859 texts distributed among six great areas grouped in three schools of knowledge, totalling 40,246,492 words.

Table 6.9 presents the statistical information on the corpus contents. As we can see, the two subcorpora, comprising texts written in Brazilian Portuguese and European Portuguese respectively, are nearly the same size. Similar balance in size of both varieties can also be found in great areas and schools, whereas the balance between great areas and schools is very much in favour of Humanities.

Table 6.9 Statistical information on CoPEP

		Whole corpus	***Brazilian Portuguese***	***European Portuguese***
	Texts	9,859	3,795	6,064
	Words	40,246,492	20,149,980	20,096,511
	Tokens *(also for data below)*	48,548,527	24,306,513	24,242,014
Humanities		**30,955,740**	**15,458,157**	**15,497,583**
	Human Sciences	25,570,856	12,763,567	12,807,289
	Applied Social Sciences	5,384,884	2,694,590	2,690,294
Life Sciences		**16,118,303**	**8,099,937**	**8,018,366**
	Health Sciences	13,515,763	6,787,116	6,728,647
	Agricultural Sciences	2,602,540	1,312,821	1,289,719
Ex-Tech-Multi Sciences		**1,474,484**	**748,419**	**726,065**
	Exact-Earth Sciences	793,877	400,040	393,837
	Engineering	680,607	348,379	332,228

It should be noted that the imbalance between schools and great areas is the reflection of the distribution of published documents per great area in each national collection on the SciELO platform, as shown in Figure 6-7[176] and Figure 6-8.[177]

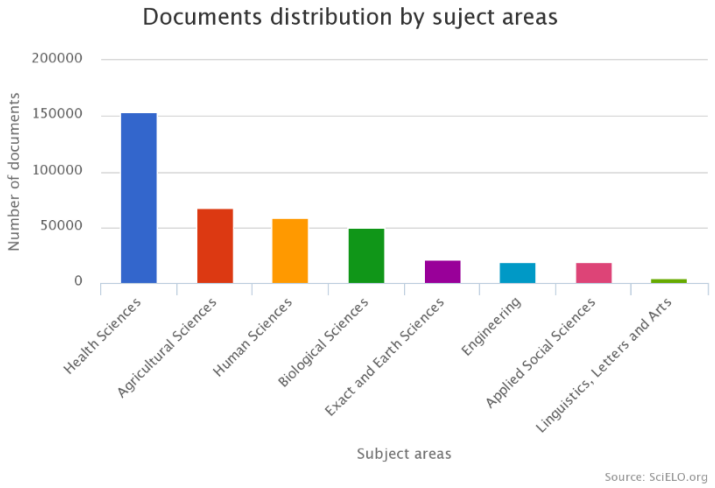


Figure 6-7 Documents distribution by great areas in SciELO-Br

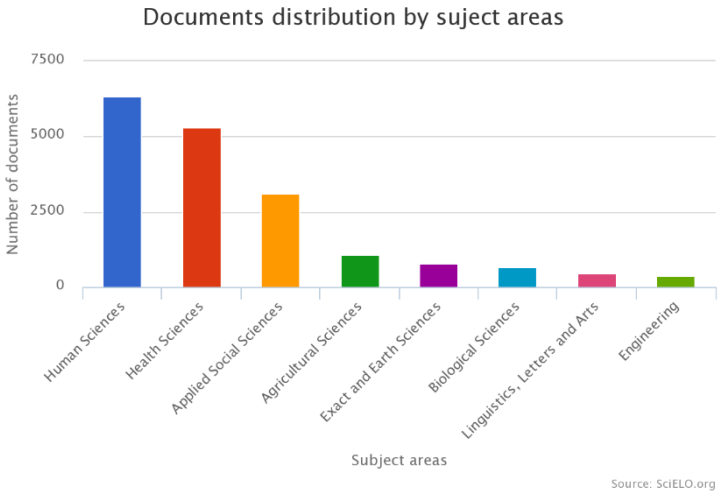


Figure 6-8 Documents distribution by great areas in SciELO-Pt

[176] Scielo Brazil Analytics http://analytics.scielo.org/w/publication/article?collection=scl (accessed 02 February 2017)

"Suject" [SIC]

[177] Scielo Portugal Analytics: http://analytics.scielo.org/w/publication/article?collection=prt (accessed 02 February 2017)

"Suject" [SIC]

Although CoPEP is relatively small according to modern day corpus standards, its size makes it ideal for evaluation and tool development, which are usually done on sample corpora with sizes varying from 50 million to 100 million tokens, and sometimes even smaller than that. CoPEP was used for all the testing in the process of setting up automatic data extraction (Chapter 9), including sketch grammar (Chapter 7) and GDEX development (Chapter 8). The corpus also provided lexical content that served as a basis for candidate headword list definitions (Chapter 10) and for compiling pilot-entries (Chapter 11).

To that end, CoPEP was first uploaded to the Sketch Engine, where it was tokenised, lemmatised and tagged with the Freeling v3 tagger. As tools development began, it became apparent that adjustments in the corpus and the corpus tool were required due to particular characteristics of CoPEP and lack of suitable resources to deal with them. The reasons for such a requirement and a description of the corpus and corpus tool post-processing will be given in the next section.

6.4 Post-processing CoPEP

As shown above, CoPEP is a multinational corpus comprising two language varieties (BP and EP). In the case of the Portuguese language of the 20th and 21st centuries, this means the coexistence of various spelling norms in CoPEP, posing a serious challenge for lexicographic work as computational resources do not adequately attend to this diverse scenario.

Particularly significant for the context of my research is the fact that, due to this scenario, CoPEP encompasses texts following four different spellings - FO43 (Br), AO45 (Pt), AO90 (Br) and AO90 (Pt) (see Chapter 1). At this point, it should be highlighted that unification of spelling rules put forward by AO90 did not overrule existing lexical traditions, so there is still variation between BP and EP. Table 6.10 briefly illustrates the linguistic situation in CoPEP:

Table 6.10 Different spelling norms in CoPEP

	FO43	AO45	AO90	AO90
	BP	**EP**	**BP**	**EP**
frequent	*freqüente*	*frequente*	*frequente*	*frequente*
co-author	*co-autor*	*co-autor*	*coautor*	*coautor*
anti-inflammatory	*anti-inflamatório*	*anti-inflamatório*	*anti-inflamatório*	*anti-inflamatório*
characteristic	*característica*	*característica*	*característica*	*caraterística/ característica*
act	*ato*	*acto*	*ato*	*ato*
infection	*infecção*	*infecção*	*infecção*	*infeção*
anonymous	*anônimo*	*anónimo*	*anônimo*	*anónimo*
register	*registro*	*registo*	*registro*	*registo*
great	*ótimo*	*óptimo*	*ótimo*	*ótimo*

Moreover, there are cases in which AO90 admits two orthographic forms. One example can be seen with the example of the double spelling of 'characteristic'in EP: with 'c' – *característica*; without 'c' – *caraterística*. However, in the event of doubt, the norm for EP spelling suggests the word with 'c'.

It is widely known that the coexistence of multiple spelling norms and multiple language varieties of Portuguese in a corpus is a challenge for NLP resource developers. The usual approach involves processing and creating tools and resources for EP and BP separately. With regards to tool and resource revision, adaptation, and creation as a result of the recent implementation of AO90, attempts have been made to tackle this issue both with a focus on only one language variety (e.g., for BP, Calcia, 2015) and various varieties (e.g. Almeida et al. 2013; Garcia, Gamallo, Gayo, & Cruz, 2014).

Given that this challenge has not been fully addressed yet, computational processing of CoPEP presented some problems, which will be described below together with the provisional workarounds employed for the development of this PhD research. Attention will be drawn to the limitations resulting from such problems and the decisions made in order to undertake the present project regardless.

6.4.1 Problems with annotation of CoPEP

As previously mentioned (Chapter 5), Freeling v3 is the default tokeniser, lemmatiser and POS-tagging program for corpora of Portuguese in the Sketch Engine. While devising Sketch Grammar for academic Portuguese (Chapter 7), I noticed some issues with the annotation of CoPEP.

Whereas some of the problems referred to limitations of the tool and required further enhancement (which is outside the scope of this thesis), others resulted from language variation within CoPEP[178] and apparently could be minimized through some post-processing. Table 6.11 below shows two examples:

Table 6.11 Examples of annotation problems in CoPEP

Tool limitation	Language variation
Cinco POS-tag: verb lemma: cincar	trabalho_N suj de V 18,994 29.53 equipar + 585 9.50 trabalho em equipe

Particularly relevant for the lexical analysis of lemmas for writing entries for DOPU is the kind of problem regarding language variety. In the case above, the grammatical relation is noun subject of verb, where the keyword is *trabalho* ('work', noun) and the collocate is *equipar* ('to equip', verb). Nevertheless, the LCM indicates a problem as it reads *trabalho em equipe* (lit. 'work in team'), that is, a different grammatical relation, noun preposition noun. It is immediately apparent that *equipe*, which is the BP spelling of 'team', had not been identified as such (a noun), but as the third person singular of the present tense subjunctive mood of the verb *equipar*. This was confirmed when concordance lines were opened and tag and lemma were displayed, as can be seen in Figure 6-9.

Figure 6-9 Part of a concordance line in the Sketch Engine

[178] In this sense, the employment of alternative taggers would not result in better annotation as a thorough search for other solutions has shown that the vast majority of free POS-taggers are language-variety specific, whereas adaptation to AO90 is still under development.

Another examination also suggested that *equipe* as a noun was not part of Freeling's dictionary: a search for *equipe*, attribute: noun, in word sketch yielded only 56 occurrences when this is a high-frequency lemma (275.99 per million in the Portuguese Web 2011 (ptTenTen11, Palavras parsed corpus). A search for the EP spelling, *equipa*, resulted in 4,014 (82.68 per million) occurrences, indicating that Freeling v3 was created for EP.

Further examination of word sketches and concordances of other BP variants in the Sketch Engine suggested that not only did Freeling v3 cover EP only but also that the norm followed was AO45, and not AO90. For instance, a search for the lemma *refletir* ('to reflect', verb, BP FO43 and AO90; EP AO90) only found the word form *refletir*, meaning that conjugated forms were not found. Verification of the lemma assigned to the keyword indicated *refletir*, but the unusual lack of any conjugated form in the results led me to search for conjugated forms of *refletir* (e.g. *refletiu*). Unexpectedly, the lemma assigned to the conjugated forms was not *refletir*, but rather the form of the conjugated verb (I was later informed by the Sketch Engine team that when a word is not in the dictionary, the searched form is assigned as lemma). These results indicated the lack of the BP variant in Freeling's dictionary. A final search, now for the EP variant (*reflectir*, AO45) yielded many occurrences, with conjugated forms correctly assigned to lemma *reflectir*. This whole examination confirmed the inexistence of the BP variety in Freeling v3 and no modification to AO90.

6.4.2 Corpus annotation workaround

The previous investigations revealed that Freeling v3 does not comply with the demands put forward by the characteristics of CoPEP as it only covers EP AO45, lacking FO43 (BP) and AO90 (BP and EP). It should be noted that a thorough search for alternative taggers[179] has shown that the vast majority of free POS-taggers are

[179] It is known that PALAVRAS (Bick, 2000), which performs morphological tagging and syntactic parsing, covers both varieties and has very high-performance results. Nevertheless, there are certain aspects that, taken together, led me to opt not to use it in this thesis. First, it is a paid resource. Although it is possible to negotiate prices for academic use, it should be noted that this is a project with plans to be implemented, thus it is wise to conceptualize the dictionary in a way that potential developments, for instance, the use of a larger corpus, are not hampered due to corpus annotation limitations. Moreover, my research adopts the ultimate methodology of data acquisition for lexicographic work, which uses POS-tagged, rather than parsed, corpora. It would be very interesting, though, to experiment the procedure with a parsed-based sketch grammar and compare the two processes. Finally, it has been reported that the

language-variety specific, whereas adaptation of dictionaries to AO90 is still under development.

Given these conditions, a tentative workaround for using CoPEP and Freeling was found to be the best possible solution for the moment. It included normalization of CoPEP to AO90 orthography, thus reducing variation resulting from the presence of multiple norms, and adapting Freeling dictionaries in the Sketch Engine.

As explained above, previous to the agreement, each variety followed different norms, namely, FO43 in Brazil and AO45 in Portugal. Thus, the AO90 unification of BP and EP spelling rules involved different processes of orthographic transformation for each variety. For instance, AO90 bans accentuation of letter *u* in the context *[qg]ü[ei]*. As *ü* is not a norm of AO45 (EP), this new standard applies only to BP, whose texts must undertake a conversion rule for elimination of the diaresis (e.g., pre-AO90: *freqüente*; post-AO90: *frequente*). Conversely, AO90 defines that the letter *c* must be eliminated when it is not pronounced. This norm applies to EP but not to BP (e.g. pre-AO90: *infecção* (BP and EP); post-AO90: *infecção* (BP), *infeção* (EP)). Consequently, each subcorpus had to be converted separately.

Thus, Lince (Ferreira, Lourinho, & Correia, 2012), a free stand-alone application for conversion of texts to the current orthography, was run on each variety, resulting in a normalised corpus named CoPEP_ AO90.

However, this action solved only part of the problem, as Freeling still lacked identification of BP variants and variants resulting from AO90. After exposing the problem to the Sketch Engine team, a solution was found to resort to Freeling v2, which was indicated to be a model for BP, taking its dictionary and merging the two Freeling dictionaries.

The process thus involved running Lince in each dictionary, that is, the BP dictionary from Freeling v2 and the EP dictionary from Freeling v3, then merging them into one. It should be stressed that the Freeling affixation rules file was analysed, revealing the reform had no effect.

corpus needs extensive processing after parsing in order to make dictionary examples searchable for lexicographers, demanding a great deal of time and specialized personnel. Due to all of these reasons, the best solution was to maintain the Sketch Engine's default tagger.

This whole procedure – CoPEP post-processing and Freeling dictionaries processing and merging – considerably reduced language variation, allowing a more homogenous corpus analysis, producing more realistic statistical computations and providing better possibilities for finding good dictionary examples. Nevertheless, there are still some limitations and further improvement is required.

6.5 Summary

Post-processing of CoPEP and the Freeling tagger was a workaround measure to allow development of my research in a rigorous and (as much as possible) accurate manner, bearing in mind the conditions of production of this thesis.

CoPEP_AO90 was used for developing the final versions of Sketch Grammar and GDEX as well as for the second procedure of automatic extraction of data from the corpus and import into iLex, meaning that DOPU was designed (i.e. the macrostructure was defined and the pilot-entry was compiled) using CoPEP_AO90 as a source of lexical information. This was only possible due to the processing of Freeling as well.

Chapter 7 Sketch grammar for academic Portuguese

As explained in Chapter 4, the procedure of automatic extraction of data from the corpus is grounded on word sketches (see Chapter 5), which is an exclusive function of the Sketch Engine tool (Kilgarriff et al., 2004) for lexical profiling. For building word sketches, the requirements are: a POS-tagged corpus and a sketch grammar. While the compilation of a new corpus was accounted for in the previous chapter (6), this present chapter sets out to report on the development of a new sketch grammar especially built for CoPEP.[180]

A brief description of the structure of sketch grammar is given in 7.1. Then I move on to a review of existing sketch grammars for Portuguese in order to evaluate their suitability for the purposes of my research (7.2). My findings revealed the need to develop a new sketch grammar, which is described in detail in section 7.3.

7.1 Sketch grammar

Sketch grammar is a file with grammatical relations, or gramrels, and processing directives for the Sketch Engine system[181] to compute different types of relations through statistics calculations.[182] The data obtained with these computations then form the basis of the word sketch feature in the Sketch Engine, and relatedly, the Thesaurus and Sketch Diff features (see Chapter 5).

Sketch grammars devised for POS-tagged corpora use regular expressions over POS-tags to find matches for grammatical relations. Queries are written in Corpus Query Language (CQL), with attribute-values names following the tagset originally used for corpus tagging. Gramrel names are preceded by the equal (=) sign and a brief description of the grammatical relation searched for.

*UNARY is the directive used for one-labelled queries that match unary relations, for example, verbs used in the reflexive form in the BNC corpus:

[180] A great majority of the information given in this chapter is also available in Kuhn and Kosem (2016).

[181] https://www.sketchengine.co.uk/documentation/writing-sketch-grammar/

[182] https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/

```
*UNARY
=as reflexive
1: VERB [tag="PP" & word = ".*sel[fv].*"]
```

In a word sketch search for the verb *cut*, this query finds concordances like *if your child cut himself*, *he cut himself a slice of bread*, and *she cut herself off from us*, indicating different constructions and meanings of the verb *cut* as a reflexive.

For binary relations, two labels are set, "1:" for the keyword and "2:" for the collocate. In a SYMMETRIC relation, both positions have the same tag attribute. For example, a search for *buy* will display collocations like *buy and sell* and *buy or rent* for gramrel =and_or.

*DUAL directive is used to tell the system to swap labels so that the second gramrel can be matched by means of the same query, only with inverted positions of labels. Forward slash (/) is the sign for separating different gramrels in the DUAL directive. On this example for English, two binary relations involving verbs and objects are translated into the following gramrels and query:

```
*DUAL
= objects of verb X/verbs with X as object
1: [tag= "V.*] 2: [tag= "N.*"]
```

Taking *buy* as the keyword, this query captures results like *buy house*, *buy shares*, and *buy ticket* for the first gramrel. The same query, with swapped labels, matches these verbs for the second gramrel when *house* is the object: *build (house)*, *buy (house)*, and *sell (house)*.

*TRINARY relations include a third label. One example from the sketch grammar for the BNC corpus is the gramrels ="%w" %(3.lemma) .../... %(3.lemma) "%w", which find prepositional phrases. A search for *buy* as keyword yields results like *buy in bulk*, *buy over the counter*, and *bought with money*.

7.2 Sketch grammars for Portuguese

There are currently four different sketch grammars for Portuguese available in the Sketch Engine. The first one is the Sketch Grammar for Portuguese, FreeLing tagset

version 1.1[183] (henceforth FreelingSkG), the Sketch Engine's default sketch grammar for all corpora of Portuguese tagged with the Freeling tagger. In fact, this is the sketch grammar devised for the Spanish TenTen corpus, which uses the same tagger, meaning that the queries were originally meant for capturing grammatical relations in Spanish, not Portuguese.

The second sketch grammar for Portuguese is the one used for the PtTenTen [2011, Palavras parsed] corpus[184] (henceforth PalavrasSkG). This sketch grammar was devised especially for the compilation of the Oxford Portuguese Dictionary (2015), a corpus-based dictionary of Portuguese-English/English-Portuguese, in which both the Brazilian and the European varieties are accounted for. The Sketch Engine computes word sketches automatically from these kinds of parsers, so PalavrasSkG does not contain CQL-written queries, but only the names of the grammatical relations.

Two other sketch grammars for Portuguese were available at the time of research: the Compatible Portuguese Sketch Grammar definition (henceforth AraneaSkG), devised for the Aranea (Web) Corpora Family (Benko, 2014a) and the Portuguese word sketches (Linguateca parsed data) version 1 (henceforth LinguatecaSkG), written for the Cetenfolha, Cetempublico corpus.[185]

The Araneum Portugallicum Maius [2015] corpus[186] is part of a family of comparable web corpora to be used for contrastive linguistics and bilingual lexicographic projects (Benko, 2014a; 2014b). Word sketch results of different languages are comparable due to a set of compatible sketch grammars, which are not syntactically-based like other sketch grammars available in the Sketch Engine, but rather collocationally-based, and whose purpose is to provide uniform results among different languages, i.e. it is intended to be language non-specific.

---

[183] This is the version I used in my research. The latest version is 1.1.1: https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/pttenten11_freeling_v3_1

[184] This corpus is part of the TenTen corpus family (Jakubíček et al., 2013), which consists of very large corpora (billions of tokens) compiled by web crawling and processed by the Sketch Engine team. The Palavras-parsed version of this corpus contains 2,757,635,105 words and is composed of texts from .pt and .br urls. As previously mentioned, this corpus is no longer accessible at the time of writing.

[185] This corpus has been renamed to Newspapers in Portuguese (CetemPúblico CetenFolha) https://the.sketchengine.co.uk/bonito/corpus/first_form?corpname=preloaded/portuguese.

[186] It contains 862,134,902 words. As a web corpus, it was compiled by web crawling and prepared by Benko (2014a, 2014b). See http://ucts.uniba.sk/aranea_about/index.html.

Cetenfolha, Cetempublico contains over 56 million words and is a combination of parts of two corpora comprising extracts of texts from two widely read daily newspapers from Brazil and Portugal: Folha de São Paulo (Brazil) and Público (Portugal). CETENFolha (24 million words, Brazilian Portuguese) and CETEMPublico (180 million words, European Portuguese) were compiled by the Linguateca[187] project and, like all the other corpora available in this resource centre, processed by the PALAVRAS dependency parser (Bick, 2000). As mentioned above, the Sketch Engine computes word sketches automatically from these kinds of parsers. Nevertheless, this corpus has a sketch grammar with CQL-written queries where both regular expressions over POS-tags pattern matching and Constraint Grammar tags are used.

Preliminary tests of the performance of the default sketch grammar for Freeling-tagged corpora in the Sketch Engine, using a sample corpus of 5 million words (henceforth the sample-5mil corpus), comprising files from Scielo.pt, revealed several problems, as will be shown below.

Such poor performance led me to the conclusion that the default sketch grammar could not be directly applied to CoPEP, hence a new sketch grammar had to be developed for my corpus. The first step of this process was to evaluate, besides the default grammar, other existing sketch grammars for Portuguese, in order to determine whether they, or their parts, could be used for my project.

7.2.1 Evaluation of FreelingSkG and PalavrasSkG[188]

Previous analysis of vertical files of the Freeling-tagged version of the sample-5mil corpus indicated that, in the majority of cases, words with capital letters were tagged as proper nouns, independently of their actual category. Word sketches of a series of lemmas indicated that participial adjectives were never matched in gramrels with adjectives. A quick inspection revealed that this was due to the fact that participle forms were tagged as verbs by Freeling v3. In addition, some word sketches returned empty results, while others yielded wrong results. These findings indicated that

[187] As presented in Chapter 3, Linguateca is a platform with language resources for Portuguese. For more information, see http://www.linguateca.pt/.

[188] It is out of the scope of this paper to reproduce the process of evaluation and their results step by step.

FreelingSkG, which is the official sketch grammar available for Freeling-tagged corpora in the Sketch Engine and, as such, would be applied to CoPEP, was problematic.

It was necessary to compare FreelingSkG with another sketch grammar, so that there would be standards for deciding which gramrels could be maintained in full, which ones would need to be revised, and identify any missing gramrels for which completely new queries would be required. PalavrasSkG was chosen to be the contrasting sketch grammar because it had been devised especially for the compilation of a dictionary of Brazilian and European Portuguese varieties.

It should be mentioned that AraneaSkG and LinguatecaSkG were also used, but only for consultation purposes at a later stage, when developing the new sketch grammar; the queries from the two sketch grammars were compared to the ones being developed in order to provide input or better alternatives.

The evaluation focused on the coverage and accuracy of queries for different gramrels. I used a sample of lemmas from the sample-5mil corpus belonging to one of the four word classes: adjective, adverb, verb, and noun. The evaluation consisted of the following steps:

a. Take one word class (e.g. adjective);
b. List all the gramrels that include the selected word class as the keyword (%w) (e.g. `=N mod adj_%w`);
c. Take the top lemma of the selected word class from the sample-5mil corpus lemma frequency list;
d. Make a word sketch for that lemma in each corpus (the Portuguese Web 2011 (ptTenTen11, Freeling v3) corpus and the ptTenTen [2011, Palavras parsed] corpus) and analyse the results:
    I. Evaluate the results presented for all the identified gramrels by analysing samples (up to 250 concordances) of three collocates (one from the top, one from the middle, and one from the bottom of the list of the first 25 collocates ordered by salience):
        - Verify if the results are valid for the gramrel in question;
        - Investigate clues indicating which queries were evaluated (for PalavrasSkG only).

II. Take note of any errors;

e. Point out possible missing grammatical relations for the selected word class, and make note of potential queries for those relations.

f. Repeat the procedure for other word classes.

7.2.1.1 Evaluation of FreelingSkG

The evaluation of FreelingSkG was conducted on the Portuguese Web 2011 (ptTenTen11, Freeling v3) corpus.[189] As previously mentioned, preliminary tests of the Freeling tagger with the sample-5mil corpus indicated that a) in the majority of cases, words with capital letters are tagged as proper nouns, independently of their actual category; b) participle forms are always tagged as verbs; c) some word sketches return empty results; and d) some word sketches return incorrect results.

With regards to FreelingSkG, certain errors were expected due to the fact that the sketch grammar was originally written for Spanish rather than Portuguese. One example is the symmetric gramrel =and_or, which returns wrong results. This is due to the use of words *y* and *o* in the gramrel, which are the Spanish equivalents of the English words 'and' and 'or', respectively. For Portuguese, the words *e* and *ou* should be used.

The evaluation of word sketches of selected sample lemmas not only contributed to enhancing our understanding of FreelingSkG, but has also revealed some additional issues regarding corpus tagging, besides the two previously mentioned. While some gramrels did not prove problematic, but rather incomplete, others exhibited more severe problems. All these aspects were carefully recorded in the next phase, i.e. writing the sketch grammar for CoPEP. Here I point out significant findings referring to the problematic parts.

Evaluation revealed that gramrels `=adj_complement` and `=predicate`, whose queries display semiauxilary verb *ser* ('to be') (tag =VS) in position 1, were never displayed in the output panel of word sketches of the selected lemmas. In order to examine the problem further, CQL searches of the queries were performed on the corpus, always returning empty results. A detailed analysis revealed that, although in the tagset (see Appendix C) VS stands for semiauxiliary verb (verb 'to be'), there are no tokens

[189] As previously mentioned, the PtTenTen corpus is part of the TenTen web corpora. This version contains 3,900,501,097 words and was tagged with Freeling v3.

annotated with that tag in the sample corpus (nor in CoPEP). The verb *ser* is tagged with VM, i.e. as a main verb. This explained why the word sketch output for those gramrels was empty.

Wrong results due to tagging errors were spotted when the examination of word sketch outputs of several lemmas showed a discrepancy between the word class of the collocate(s) and the word class defined in the gramrel. For example, it was unexpected to see verb+noun collocations such as *apresentar olhar* and *apresentar caráter* displayed in the table of results of the gramrel `=object_inf` (verb as keyword and verb-infinitive as collocate) for the keyword *apresentar* ('to present/show'). A close examination of the tags of the collocates revealed these nouns had been tagged as verbs. That indicates that the program probably interpreted the *-ar* and *-er* endings in words such as *olhar* ('look') and *caráter* ('character') as markers of the infinitive form of verbs belonging to first and second conjugation respectively.[190] Many other cases of inconsistency between gramrels and their results were found, and different types of tagging errors recorded.

In addition to the accuracy of FreelingSkG being compromised due to Spanish framed definitions and tagging errors, examination showed that FreelingSkG is also rather limited in terms of query coverage. Some noticeable examples are: no gramrels for adverbs as keywords; no gramrel for the pair adjective-noun when adjective is prenominal; and gramrels with adjectives do not include participial adjectives.

In addition, tagging errors have proven to reduce grammatical accuracy of queries, indicating that attention should be paid to tagging issues when preparing the sketch grammar for CoPEP, or in fact any corpora tagged with Freeling. These findings suggested that FreelingSkG could not be employed for the analysis of CoPEP without considerable improvement to its queries.

7.2.1.2 Evaluation of PalavrasSkG

The evaluation of PalavrasSkG was conducted on the ptTenTen [2011, Palavras parsed] corpus, which is dependency parsed by PALAVRAS (Bick, 2000). Since the Sketch Engine computes word sketches automatically from the parsing output, the

[190] In fact, in the case of *olhar*, identification of *-ar* as a marker of first conjugation verbs in the infinitive form is correct. Nevertheless, this infinitive verb is used here as a noun, so the tagger misidentified the POS-tag.

sketch grammar for this corpus is composed of a list of gramrels without queries. As a result, the evaluation of this sketch grammar also involved investigating clues to components and structures of queries for different gramrels.

The analysis of coverage of gramrels and accuracy of queries revealed a number of interesting points, which are presented here with illustrative examples:

a) Dependency relations (deprels) annotation allows the capture of collocations with a very large span window between a keyword and a collocate. For instance, the collocation *paciente apresentar* ('patient', noun; 'to show', verb) for the gramrel =N subj_of %w_V was captured in a sentence despite a 15-token-long relative clause between the keyword and the collocate, as shown in (1).

(1) Alguns autores relatam que **pacientes** que escovam suas superfícies radiculares com estimuladores dental (uma forma de estimulação mecânica) *apresentam*/V/ dentina lisa, dura e sem sensibilidade.

b) Deprels annotation allows matching of inverted constructions. For the case of personal verbal passive constructions, i.e. agent of the passive + verb 'to be' + main verb in the participle form, simple position-based queries, which follow the canonical subject + verb + object order, detect the first element as a subject of the main verb, while it is, in fact, its object. Thus, with deprels annotation, verb-object collocations are captured, regardless of the keyword and noun positions in the sentence. This is the case of concordance (2).

(2) onde as **propostas** *definidas nas plenárias regionais pelos sindicatos filiados serão* apresentadas/apresentar/V/ e submetidas à votação na Plenária Estatutária , que acontecerá, no dia 27 de junho, às 10h na sede do Sindicato dos Metalúrgicos do Abc

c) Although there were only 13 gramrels in PalavrasSkG, the deprel attribute view option showed annotation of many more relations than stated. For instance, when analysing the word sketch output for the verb-object pair relations, with verb as the keyword, some other deprels involved in matching such a collocation were:

```
=%w_V obj N
*UNIMAP object
#DEPREL
# ADJ _por_ %w_V/%w_ADJ _por_ V
# NUM subj_of %w_V/%w_NUM subj_of V
# SPEC subj_of %w_V/%w_SPEC subj_of V
# V comp %w_V/%w_V comp V; ,%w_V obj N/V obj %w_N
```

d) Identification of tagging errors when adjectives were in prenominal position. Such constructions are marked in Portuguese and were not explicitly covered by PalavrasSkG. Thus, CQL searches of sample lemmas (adjectives) followed by a noun were performed to confirm this missing gramrel coverage. To my surprise, concordances seemed to contain good collocations, i.e. the sample adjective lemmas followed by correct noun collocates, but a detailed inspection showed that those were false positives, since the original collocation adjective+noun was only matched due to wrong tagging of both the keywords and the collocates. In other words, adjective lemmas had been tagged as nouns and noun lemmas as adjectives.

(3) 

e) Identification of bad collocations. Attributing wrong collocates to the keyword seems to be due to parsing errors. In the concordance (4), there are two clauses linked by the conjunction *e* ('and'). However, the system matched the subject of the first clause (*relator*, 'rapporteur') with the main verb (*apresentar*, 'give') of the second clause. In fact, *relator* is the subject of *ligou*, and *nos* is the first person plural pronoun in its indirect object form. In the second clause, the subject pronoun *nós* is omitted because the inflection of the verb (*vamos*) indicates person (first) and number (plural).

(4) "O **relator** *nos ligou e vamos* *apresentar* nossas opiniões para ele", disse Maia.

Although PalavrasSkG could not be applied to CoPEP due to its parsing (and not POS-tagging) annotation, the findings of this evaluation have had significant implications for the understanding of this sketch grammar. First, several possible grammatical relations in Portuguese were recorded, with a special note on categories and occurrences of items between keywords and collocates. Second, dependency relation annotation proved to be especially valuable for finding relations between items that have several intermediary words, and in inverted constructions. Finally, occasional failure in capturing collocations helped me record different errors for future reference. Overall, it was found that PalavrasSkG covers an extensive set of grammatical relations and contains very accurate queries.

The evaluation of FreelingSKG and PalavrasSkG has revealed advantages and shortcomings of both sketch grammars, as well as several issues of the Freeling tagger, which has been used for tagging CoPEP. Nonetheless, the overall conclusion was that neither of these sketch grammars could be used for my purposes, but rather that a completely new sketch grammar for academic Portuguese would need to be developed; this new sketch grammar could, however, still utilise some of the good gramrel queries, or parts of them, from the sketch grammars evaluated above.

7.3 Devising a new sketch grammar for academic Portuguese

Devising the sketch grammar for academic Portuguese (henceforth AcadPortSkG) consisted of two phases: writing gramrel queries and evaluation. The first phase was grounded on a trial and error method, where queries were written and tested many times until satisfactory results were reached. This process was not only laborious but also time-consuming: for every new attempt, the corpus had to be recompiled in the Sketch Engine. To speed up corpus recompilation and analysis, a sample 1-million-word corpus was used instead of the entire corpus. The second phase took place once the sketch grammar[191] was deemed sufficient for the purpose of experimenting with automatic extraction of data from CoPEP. After recompiling the

[191] The current version is AcadPortSkG_v3. At the time of the experiment, version 7.1 was used. For every new version of AcadPortSkG, the results of changes were evaluated in the same manner as described here.

corpus with the new sketch grammar, the evaluation of word sketches of sample lemmas was conducted.

All this work came down to a sketch grammar with symmetric (1), unary (3), dual (14), and trinary (2) grammatical relations covering attributive (pre- and postpositional) and predicative adjectives; nouns as predicative complement, subjects, and objects of verbs (unmarked order); prepositional phrases; infinitive as verb/noun/adjective complement; impersonal and personal verbal passive constructions; impersonal constructions with *se*; verbs followed by *que*-clauses (subordinate clauses); verbs with gerund as complement, and adverb-verb and adverb-adjective pairs. Figure 7-1 shows a partial view of the word sketch results for the lemma *estudo* ('study', noun).

estudo *(noun)*
CoPEP freq = 81,688 (1,682.60 per million)

sintagma preposicional			estudo_N suj de V			estudo_N mod por Adj-Part		
	70,299			32,156	0.39		19,779	0.24
...deestudo	22,327	0.27	mostrar +	1,030	9.42	realizar +	2,262	11.27
...emestudo	15,212	0.19	demonstrar +	809	9.36	estudo realizado		
estudo de N	14,147	0.17	ir +	4,092	9.19	recente +	771	9.98
estudo sobre N	3,627	0.04	estudo foi			estudo recente		
estudo em N	3,601	0.04	ter +	2,518	9.01	anterior +	666	9.57
...paraestudo	2,132	0.03	revelar +	648	8.85	estudos anteriores		
estudo com N	1,861	0.02	apontar +	498	8.71	empírico +	540	9.48
...aestudo	1,826	0.02	estudos apontam			estudos empíricos		
...porestudo	1,289	0.02	avaliar +	410	8.57	comparativo +	355	9.07
...comestudo	1,045	0.01	permitir +	610	8.38	estudo comparativo		
estudo por N	1,018	0.01	utilizar +	386	8.33	publicar +	364	9.03

Figure 7-1 Partial results of word sketch for *estudo*

The results of trinary relations open on a separate page, allowing for a search of each individual relation. Partly visible in Figure 7-1, for instance, are 35 gramrels for *estudo* followed or preceded by a prepositional phrase (the gramrel column titled *sintagma preposicional*), each of them with their own column of collocates. For example, of the 22,327 occurrences of **…de estudo,** the collocation *resultado do estudo* ('result of study') represents 1,521 occurences; similarly, of 14,147 occurrences of **estudo de N**, the collocation *estudo de caso* ('case study') represents 1,317 occurrences.

To give some indication of the number of gramrels per lemma, I provide the data for the top five frequent lemmas per word class (Table 7.1). We can see that the generalization of number of gramrels per word class can only be made to some extent. It

is possible to affirm which gramrels do not take certain word classes as keywords, namely, the three unary relations: no adverbs and nouns; both types of trinary relations: no adverbs; and prepositional phrase trinary relations: no adjectives. Besides that, numbers vary according to the characteristics of the keyword in question.

Table 7.1Numbers of gramrels for the five most frequent lemmas in each word class

		Symmetric	Unary	Dual	Trinary	total
Top 5 nouns	estudo	1		10	Prep.phrase: 36 Prep+inf: 7	54
	relação	1		10	Prep.phrase: 36 Prep+inf: 6	53
	forma	1		10	Prep.phrase: 35 Prep+inf: 7	53
	ano	1		10	Prep.phrase: 31 Prep+inf: 9	51
	trabalho	1		10	Prep.phrase: 37 Prep+inf: 8	56
Top 5 verbs	ser	1	3	7	Prep.phrase: 23 Prep+inf: 12	46
	ir	1	3	8	Prep.phrase: 24 Prep+inf: 11	47
	ter	1	3	11	Prep.phrase: 22 Prep+inf: 9	46
	poder	1	2	6	Prep.phrase: 17 Prep+inf: 9	35
	estar	1	3	8	Prep.phrase: 22 Prep+inf: 7	41
Top 5 adjectives	social	1	1	5	Prep+inf: 8	15
	maior	1	1	5	Prep+inf: 6	13
	novo	1	1	5	Prep+inf: 7	14
	político	1	1	5	Prep+inf: 7	14
	primeiro	1	1	5	Prep+inf: 4	11
Top 5 adverbs	não	1		4		5
	mais	1		4		5
	também	1		4		5
	assim	1		4		5
	ainda	1		4		5

In the following section, I present the procedure of writing AcadPortSkG and its evaluation.

7.3.1 Phase 1: writing

The method of writing AcadPortSkG was as follows:

1. Select part-of-speech items (noun, adjective, verb, adverb);
2. Determine grammatical relations among them;
3. Name those gramrels;
4. Define directives to find gramrels;
5. Write queries for gramrel matching in the sample corpus:
    a. Use CQL concordance search in the sample corpus to verify regex matching;
    b. If regex query works, write it in the sketch grammar file;
6. Recompile the sample corpus after each change to the sketch grammar;
7. Verify in the sample corpus if the sketch grammar works.
8. Once the sketch grammar yields satisfactory results, go to phase 2.

As an illustration of what writing a sketch grammar entails, a brief account of the process of writing queries for the grammatical relations between adjective and noun will be given below.

As mentioned earlier, a preliminary annotation test pointed out tagging of participle forms as verbs only. Nevertheless, in Portuguese, those forms can also function as adjectives. Thus, for gramrels with this category, a tag for the participle form of the verb (V.P.*) had to be added.

Firstly, a simple combination of a noun followed by adjective/verb participle (unmarked order in Portuguese) was tried out in the sketch grammar file. The name used for this gramrel was `=N mod por Adj-Part`. Here, the word sketch of lemma *social* ('social', adjective) is explored.

```
=N mod por Adj-Part
1:"N.*" 2:[tag="A.*"|tag="V.P.*"]
```

As expected, the majority of collocations found were valid, for example, *ciências sociais* ('social sciences') and *classe social* ('social class').

The marked order of adjectives in Portuguese, i.e. before nouns, is not covered by FreelingSkG. The test sketch grammar was edited with the directive *DUAL for the

two gramrels =N mod por Adj-Part/ =Adj-Part mod N. The word sketch for test lemma *estudo* ('study', noun) yielded good matches like *estudo analítico* ('analytical study') for the former gramrel, and *presente estudo* ('present study') for the latter.

After confirming that these two gramrels worked fine on the sample 1-million-word corpus, different intervening words were tried out. For that, new queries were written and searched in the sample corpus. If these queries produced good results, they would be included in the test sketch grammar, and after each such change was implemented, the sample corpus had to be recompiled in the Sketch Engine.

To sum up, the experiment involved, firstly, addition of one optional adverb, then one optional adjective following the optional adverb. Analysis of word sketches of a number of lemmas showed that these two extra optional items increased the number of good matches for the gramrel noun+adjective. For the reversed gramrel, i.e. adjective+noun, the second adjective preceding the noun returned bad matches in most cases,[192] whereas adverbs were not tested because they do not occur in this position in Portuguese (cf. Perini, 2002). Next, the number of optional intermediary adverbs and adjectives were expanded to two in the gramrel `=%w_N mod por Adj-Part`, which yielded more results (which were still good) than its initial version. Hence, collocates within a wider span were found, for example:

(5) Em *estudo* ainda não publicado **realizado** no Peru, também foi possível reproduzir os pseudopólipos com a utilização de..

In (5), the collocation *estudo realizado* was found in a construction with two adverbs (*ainda* and *não*) and one adjective (*publicado*).

Finally, an attempt to capture adjectives collocating with the head of a noun phrase composed of head noun + prepositional phrase[193] + adjective led to the inclusion

[192] The very few good matches mostly referred to some interesting combinations like *tradicionais boas relações* (lit.,'traditional good relationships') and *suposta inerente vulnerabilidade* (lit., 'supposed inherent vulnerability'), in which the adjective immediately preceding the noun and the noun seem to form a lexical unit which is qualified by the first adjective. As there was much more noise than good matches, I opted for not including intermediary items in the query for the gramrel adjective + noun.

[193] Perini (2002) uses the term *modifier* to refer to (preposed and postposed) words that modify the head of noun phrases. According to him, "a modifier can also be composed of a prepositional phrase" (Perini,

of optional intermediary prepositions and determiners. A query like that would capture good collocations like *trabalhos desenvolvidos* (lit.[194] 'works conducted') out of the structure *trabalhos de investigação desenvolvidos* (lit., 'works of research conducted'). Nevertheless, the examination of concordances revealed the presence of too much noise, which severely hindered the accuracy of collocation matching. Thus, these items were excluded.

In the end, these were the gramrels devised for capturing collocates between adjective and noun, where R stands for adverbs:

```
*DUAL
=Adj-Part mod %w_N/%w_Adj-Part mod N
2:"A.*|V.P.*" 1:"N.*"
*DUAL
=%w_N mod por Adj-Part/N mod por %w_Adj-Part
1:"N.*""R.*"{0,2}"A.*|V.P.*"{0,2}2:"A.*|V.P.*"
```

It is noteworthy that this phase of writing gramrel queries not only resulted in a new sketch grammar for academic Portuguese, but also contributed to the improvement of the overall quality of my research due to two important revelations, which, in turn, led to correction measures.

Firstly, verification of regex matching through CQL concordance searches in the sample corpus revealed 'junk' in the corpus, such as the following textual passages "[Creative_Commons_License]", "texto apenas em PDF", "abstract", "resúmen", besides email addresses, phone numbers, and texts written in languages other than Portuguese. Consequently, extra cleaning of the corpus was performed and its quality significantly enhanced.

Secondly, other types of tagging errors, in addition to the ones already listed in the previous sections, were spotted during this phase. Attempts to work around problems related to the identified tagging errors demanded unique approaches to

2002, p. 327). Prepositional phrases are contiguous with the heads when such a phrase is a classifier; in case of a second modifier (an adjective), this comes at the end of the noun phrase, as in the example above.

[194] In many cases translations are literal and not necessarily correct in order to reflect the structure/gramrel.

different gramrels, making the whole process more complex. A few examples of such workarounds are discussed in the next section.

7.3.2 Phase 2: Evaluation of AcadPortSkG on the CoPEP corpus (40 million words)

After developing AcadPortSkG on a sample corpus, I moved to its evaluation on the CoPEP corpus data. This entailed compiling the corpus in the Sketch Engine using AcadPortSkG, defining a methodology of evaluation, conducting the evaluation, and proposing workarounds for gramrels in which annotation problems seriously affected the results.

The objective of the evaluation was to verify whether the devised gramrel queries captured correct information. A sample lemma list for the evaluation was selected according to the following two criteria: frequency and diversity of characteristics of each word class. By diversity, I mean heterogeneity of word class characteristics, e.g. verbs with different valency patterns (transitive, intransitive); descriptor and classifier adjectives; adverbs of manner, degree, time; abstract and concrete nouns.

Using the Word List function in the Sketch Engine, I created a list of lemmas ordered by frequency. Different frequency bands were set considering the size of the corpus: low-frequency lemmas were considered those with a frequency between 500 and 1,000 (between 9.34 and 18.69 occurrences per million words); mid-frequency lemmas were those with a frequency between 3,000-5,000 (between 56.10 and 93.40 occurrences per million words); and words with a frequency of more than 5,000 (93.40 occurrences per million words) were considered high-frequency lemmas. Then, for nouns, adjectives and verbs, I selected 50 high-frequency lemmas, 15 mid-frequency lemmas, and 10 low-frequency lemmas, i.e. 75 lemmas per word class. For adverbs, I selected 45 lemmas (30 high-frequency, 10 mid-frequency and 5 low-frequency lemmas).

For the evaluation of the sketch grammar results, I used the following procedure:

1. Make a word sketch for one of the lemmas from the list;

2. Examine each gramrel in the word sketch, following these steps:

a) When longest-commonest match (LCM) is displayed (lines in grey under the collocates (see Figure 7-2) are the most common realization of the collocation in the corpus) check if the collocation seems good[195] (in this way, I also checked the usefulness of information in LCM);

estudo_N suj de V		
	32,156	0.39
mostrar +	1,030	9.42
demonstrar +	809	9.36
ir +	4,092	9.19
estudo foi		
ter +	2,518	9.01
revelar +	648	8.85
apontar +	498	8.71
estudos apontam		

estudo_N mod por Adj-Part		
	19,779	0.24
realizar +	2,262	11.27
estudo realizado		
recente +	771	9.98
estudo recente		
anterior +	666	9.57
estudos anteriores		
empírico +	540	9.48
estudos empíricos		

Figure 7-2 Partial word sketch results for *estudo* ('study', noun)

b) Examine the list of collocates and determine if most of them seem good;

c) examine the first 20 concordances of each of the top 25 collocates;[196]

d) if bad matches are found, examine more concordances;

3. Consider the ratio between good and bad matches and:

a) if there are many more good matches than bad ones, consider the gramrel good;

b) if there are many bad matches, take note of the errors.

Naturally, not all the steps of this procedure were followed for all the lemmas. If for a certain gramrel, good matches were found in over half of the lemmas, it meant that the query was picking up the correct information; thus, evaluation would move forward. The same goes for concordance reading. Presence of the right collocation in longest-commonest match, supported by a quick examination of the concordances, resulted in

[195] In the paper "A quantitative evaluation of word sketches" (Kilgarriff, Kovář, Krek, Srdanovic, & Tiberius, 2010), human experts (linguists and lexicographers) were asked to assess collocations and determine whether collocates were "Good; Good but wrong grammatical relation or POS-tagging error; Maybe (not striking collocate); Maybe (specialised vocabulary), and Bad" (ibid, p. 376). I followed the same categorization in my evaluation.

[196] Collocates were sorted by salience score, that is, by the strength of the collocation. Minimum collocate frequency was 3.

analysis continuing to the next collocate. However, in case of problematic collocates, more concordances would be examined.

On the one hand, the evaluation corroborated the effectiveness of a handful of gramrels; on the other, it indicated that some of them could use some improvement in order to obtain more varied results, especially due to the verification of some tagging errors, like the case of participles tagged as verbs. One example of a gramrel yielding mainly good matches was a verb followed by a preposition and another verb in its infinitive form:

```
*TRINARY
=%w_V %(3.lemma) Vinf/V %(3.lemma) %w_Vinf
1:"V.*" " 3:"SP.*" 2:"VMN.*"
```

The evaluation of word sketches revealed the query picked correct information. For instance, for the following verbs, these are some matches found:

Começar: começar a ser, ter de começar

Lembrar: deixar de lembrar, lembrar de ter

Desistir: desistir de participar, pensar em desistir

A quick read of the concordances already indicated some typical patterns, like *lembrar de ter* followed by verb in past participle (*lembrar de ter usado, visto*), thus corroborating the validity of these gramrels.

In gramrels that produced a vast majority of good results and only a few problematic collocates, the latter were disregarded. This was done because, firstly, most of them originated from tagging errors[197] with no impact on the accuracy of the query, and secondly, those cases were not predominant.

7.3.3 Improving AcadPortSkG

Nevertheless, some queries were indeed subjected to further improvement in order to improve the accuracy of their output. The main adjustment concerned finding ways to work around some corpus annotation errors. I present here two cases of

[197] For example, nouns whose forms are the same as verbs, e.g. *poder* (verb: 'can'; noun: 'power'), are tagged as verbs. Thus, the gramrels above yield occurrences like *no que se refere ao poder da Igreja Católica* ('as regards the power of the Catholic Church').

adjustment of a different nature: the first one concerning tokenization of verbs with the particle *se*, and the second one relating to lack of tagging of participles as adjectives. Additionally, I describe the process of creation of a new grammatical relation (symmetric), based on other sketch grammars of Portuguese. The reason for this new device was that the original gramrel in the sketch engine default sketch grammar had this query written with Spanish lemmas.

The particle *se* has many uses in Portuguese, thus its importance: 1. as a personal pronoun: reflexive pronoun, object; reflexive pronoun, indirect object; reflexive pronoun, object of reciprocal verbs; reflexive pronoun, object of infinitive; passive voice; unknown subject; expletive; part of verb expressing feelings, change of state, movement (Cegalla, 2008, pp. 562-563); 2. as a conjunction. It is known that those uses can be clearly determined by word sketches from dependency-parsed corpora. However, if this particle is correctly tagged, sketch grammar based on regex over POS-tagged corpora allows lexicographers to interpret its uses by analysing good concordances which reflect typical patterns.

Unfortunately, this was not the case with Freeling v.3 for Portuguese. These are the annotation errors involving *se* that have been found in CoPEP:

a) *Se* is tagged as a pronoun when it is actually a conjunction:
   Queremos saber se/se/PP3CN000 a inserção de ociosidade nessa dada seqüência pode promover uma diminuição no valor da função-objetivo.
b) *Se* is tagged as a proper noun due to capital letter;
c) Most of the time, *se* is not tokenised when postponed (thus connected to the verb by a hyphen). Instead, it is considered a unit with the verb, forming the lemma *verb+se*.

   Verbs matched are inflected for mode, tense, person, and number:

   - verb modes: indicative, subjunctive, imperative, infinitive, and gerund
   - verb tenses: present, imperfect, future, past, conditional, pluperfect[198]
   - person: 3rd person
   - number: singular, plural
   - gender: 0 (non-specified attribute; only for participle)

[198] Verbs tagged as in pluperfect tense were actually in past tense.

Examples:

Present: *deve-se* /VMIP3S0+PP3CN000/dever+se

Past: *desenvolveu-se* /VMIS3S0+PP3CN000/desenvolver+se

d) Less frequently, *se* is tokenised, lemmatised and tagged as a personal pronoun. In those cases, the hyphen is also tokenised and tagged as such (Fg). For example, the word form *escolhe-se*:

escolhe /VMIP3S0/escolher - /Fg/- *se* /PP3CN000/se

e) Since *se* is not tagged when it is part of the *verb+se* lemma, it is ignored for the analysis of the following *se*, which is tagged as a pronoun and not as a conjunction:

Verifica-se se a empresa.. ('it is verified if the company…')

verifica-se /VMIP3S0+PP3CN000/verificar+se *se* /PP3CN000/se *a* /DA0FS0/o *empresa* /NCFS000/empresa

The most significant problem to be tackled is the lack of capturing the use of *-se* when a verb is searched for in the word sketch function. This means that although a verb can occur with or without *-se*, there is no way to find such occurrences because the instances of *verb-se* are never matched.

Many different queries have been written to overcome this problem with the pronoun *se*, and all of them failed. After describing the problem to the Sketch Engine support team and showing the different workaround attempts, they proposed a reconfiguration of the Portuguese pipeline to accommodate my needs. A new corpus template - "Freeling Portuguese DEVELOPMENT" was created. Besides "lempos", "lc", and the three ordinary attributes [word, tag, lemma], three respective multi-value attributes [morphs, tags, morphemes] were added to the corpus. The attribute "morphemes" was created to account for verbs with clitics: it contains the lemma of the verb and all the pronouns (corresponding to what was joined by the "+" sign in the old pipeline). Morphological tags for the parts comprise "tags" and just the parts of the wordform separated by hyphens are "morphs", i.e. for verbs with clitics, this attribute can be the verb-stem part, the forms of the pronouns, and the suffix.

The second adjustment performed on the queries concerned the fact that the fix found for the lack of tagging participles as adjectives ended up causing a series of other

problems. The original workaround consisted in adding the tag V.P.* for adjectives, as in this gramrel:

```
*DUAL
=Adj-Part mod %w_N/%w_Adj-Part mod N
2:"A.*|V.P.*" 1:"N.*"
```

As expected, the gramrel finds good collocations like *elevado teor* (lit. 'raised level'), where the participle form is an adjective that typically collocates with the noun *teor*. Without the addition of V.P*, collocations like this one would not have been found. Nevertheless, verbs *ser*[199] ('to be'), *ter*[200] ('to have') and *haver*[201] ('to have') are primary verbs, i.e. "can function as both auxiliary and main verbs" (Biber et. al, 2015, p.104). Thus, when they precede the structure V.P*+N, in the vast majority of cases the participle form functions as a verb, not as an adjective. *Ser* makes up passive structures when followed by a participle verb form, while *ter* and *haver* followed by a participle verb form indicate a compound form with tense and aspectual functions.

For those situations, I had to come up with amendments to make sure that the gramrel matched only participle forms functioning as adjectives, not as verbs. Below, I touch upon different kinds of problems caused by adding V.P* to the query 2:"A.*|V.P.*"1:"N.*, and proposed solutions. An optional adverb ("R.*"?) was included before the participle form to allow capturing structures like são agora apresentados resultados (lit. 'are now presented results'). Each of the three verbs was dealt with separately and solutions were put together in the end to form the final query.

I) *ser* + V.P.* + noun = passive structure

This query matches, for example, *são apresentados resultados* (lit. 'are presented results'). To avoid matching participle forms as parts of passive structures, I defined that any word can precede V.P.*, except for the verb *ser*:

```
[lemma!="ser"] "R.*" ?2:"V.P.*"1:"N.*"
```

199 *Ser* as a main verb is a copular verb, i.e. it is used "to associate an attribute with the subject of the clause" (Biber et al., 2015, p. 140).

200 As a main verb, *ter* refers to the idea of possession, family connections, composition, etc.

201 As a main verb, *haver* means 'there is/there are'.

However, not all verb forms of the verb *ser* are captured by that query due to a lemmatising error in Freeling. The forms of both verbs *ser* and *ir* ('go') are the same in the simple past and in the third person plural of pluperfect. Freeling v3 tagged the forms *foi*, *foram*, *fui*, *fomos*, *foste* as *ir*, never as *ser*. Thus, another workaround was needed to fix this problem: lc!="foi|foram|fui|fomos|foste". This rule matches any item but those verb forms; lc guarantees that both upper and lower-cased words are matched.

For the cases where *ser* is a copular verb, I performed a CQL search for this lemma (and the word forms mentioned above) followed by the participle form of the verbs *elevar*, *determinar*, *limitar*, *variar*, *reconhecer*, *moderar*, which have appeared as participial adjectives when found with verb *ter*. There were only 47 occurrences in the whole 40-million-word corpus, and in only nine of them the verb *ser* was acting as a copular verb.

Despite the well-known existence of other verbs besides the ones investigated whose participle forms can also act as adjectives, the analysis of the sample verbs indicates that the occurrences of participial forms as prenominal adjectives in noun phrases in predicative function, whose linking verb is *ser*, have very low frequency when compared to the number of passive structures realised by the same word forms, i.e. *ser* acting as an auxiliary verb, participle form as a main verb, and a noun as agent of the passive.

II) compound *ter* + V:P.* + noun

In Portuguese, some compound verbs are formed by auxiliary verb *ter* followed by participle verb form. To avoid matching structures like *tem apresentado resultados* ('have presented results'), *ter* is negated from the structure:

```
[lemma!="ter"] "R.*" ?2:"V.P.*" 1:"N.*"
```

Nevertheless, this query also excludes matches where *ter* acts like a main verb, as in the sentence *A platina (Pt) tem igualmente elevada densidade* ('Platinum has equally high density'). Thus, the collocation *elevada densidade* is not found with the above-mentioned query.

This led me to analyse the corpus for occurrences of *ter* + R.*? + *elevar* (V.P*) + noun. In all 73 occurrences, *ter* was a main verb. That raised some questions: if

structures with *ter* are not to be captured, then 73 occurrences of *elevado*[202] as adjective preceding the noun will be lost; is that a considerable problem? What about other participial adjectives that will not be captured; will I miss a great deal of relevant information?

Firstly, another CQL search was performed to verify the total number of constructions *elevar* (V.P*) + noun that are not preceded by *ter*. With 6,335 occurrences for such a structure, 73 lost collocates (1.14%) are not statistically relevant. In addition, searches have shown that, out of those 73 collocates, the top three most frequent ones occurred only three times. For instance, *ter elevado risco* (lit., 'have raised risk') occurs three times in the whole corpus, while the collocation *elevado risco* occurs 202 times (only three times as the object of the verb *ter*).

Similarly, the structure *ter* + R.*? + V.P.* + noun occurs 6,399 times. Manual analysis of a random sample of 640 occurrences (10% of the total) has shown 14 occurrences of *ter* as a main verb (2.18%), eight of them with the participle form of the verb *elevar*, two with *determinado* (lit. 'determined'), and one occurrence with each of the following participial forms: *limitada* ('limited'), *variados* ('varied'), *reconhecida* (lit. 'recognised'), and *moderada* (lit. 'moderated').

In order to confirm that the exclusion of constructions where *ter* is a main verb would not affect the identification of collocations for participial adjectives, the use of those other five participial forms was investigated. The participle form of *determinar* has 9,216 occurrences when preceding nouns; 31 of these occurrences after the lemma *ter*, and 30 of these with *ter* as a main verb. The collocation *limitado* + noun occurs 247 times and only four times as the object of the verb *ter*, while *variado* + noun occurs much more often (1,034), but only five times after the verb *ter* as a main verb. The participle form of *reconhecer* preceding a noun has 246 occurrences, with five after *ter*: two of them as auxiliary verbs and three as a main verb. *Moderado* as an adjective preceding nouns occurs three times after *ter*, with 196 occurrences for that construction without *ter*.

[202] In Portuguese, lemmas of adjectives are in singular, masculine form. When in use, adjectives take on number and gender inflections.

The tests above led me to the conclusion that the option to lose some collocates for the benefit of better pattern matching seemed to be a reasonable trade-off.

III) compound *haver* + V:P.* + noun

The auxiliary verb *haver* in compound verbs has the same function as the verb *ter*. To avoid matching structures like *haviam apresentado sintomas* ('had shown symptoms'), *haver* was also negated from the structure.

Nevertheless, *haver* is used much less frequently, with 547 occurrences for *haver* + R.*? + V.P + noun. This fact alone means that any occasional loss for wrongly capturing participles used as adjectives instead of verbs would not be statistically relevant in the first place. Still, it is possible to reduce such an error by negating the verb *haver* in its plural form from the query. This is because the verb *haver* as a main verb means 'there is/there are' and is an impersonal verb, that is, it only occurs in third person singular. That rule excludes 221 occurrences of *haver*. Despite the possibility of capturing *haver* as a main verb among the remaining concordances, the number of potentially lost combinations of participial adjectives and nouns is negligible in comparison with the total amount of such combinations in the corpus (46,239 occurrences).

All the solutions proposed above had to be merged in a single query to yield the correct results. These are the gramrels and queries for matching (participial) adjective + noun:

```
*DUAL
=Adj-Part mod %w_N/%w_Adj-Part mod N
2:"A.*"1:"N.*"
[lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")&
lc!="foi|foram|fui|fomos|foste"][lemma!="ser|ter" &
!(lemma="haver"&tag="VM.*P.")&
lc!="foi|foram|fui|fomos|foste"]"R.*"?2:"V.P.*"1:"N.*"
```

Finally, a symmetric grammatical relation was created based on other existing sketch grammars, in order to overcome the fact that, in FreelingSkG, this relation used the lemmas *y*|*o*, which are actually Spanish.

In order to overcome this shortcoming, the query originally devised for the relation of coordination by conjunction in the Compatible Portuguese Sketch

Grammar[203] of the Aranea Corpora Family (Benko, 2014a) was slightly altered, its tagset converted and then adopted. In its original form, coordination is marked by the tag "conjunction", which encompasses *e*, *ou*, but also *que* (subordinating conjunction). To restrict matches to the two first, that tag was substituted by [lemma="e|ou"]. This query was chosen because it allows the capture of many word categories and tests have shown it matches more collocates than others.[204]

Table 7.2 below shows the process of adjustment of the symmetric relation. Each row refers to one sketch grammar. The third column explains the tagset used:

Table 7.2 Process of adjustment of the symmetric relation *e_ou*

Source	Queries	Tagset
Original query in default sketch grammar	*SYMMETRIC =y_o 1:"NC.*" "[AR].*"* [lemma="y|o"] "D[AI]0.*"? "[AR].*"* 2:"NC.*" 1:"V.*""[AR].*"* [lemma="y|o"]"[AR].*"* 2:"V.*" 1:"A.*" [lemma="y|o"] "R.*"? 2:"A.*"	NC: common noun A: adjective R: adverb D[AI]: determiner (definite or indefinite article) V: verb
Original query in Araneum Compatible Sketch Grammar	*SYMMETRIC 1:[atag!="Z.*"] [atag!="(Cj)|(Z.*)"]{0,2} [atag="Cj"] [atag!="(Cj)|(Z.*)"]{0,2} \ 2:[atag!="Z.*"] & 1.atag=2.atag	Cj conjunction Z punctuation
New query in the Sketch Grammar for Academic Portuguese	*SYMMETRIC =e_ou 1:[tag!="F.*"] [tag!="(C.*)|(F.*)"]{0,2} [lemma="e|ou"][tag!="(C.*)|(F.*)"]{0,2} 2:[tag!="F.*"] & 1.tag=2.tag	F: punctuation C : conjunction

As can be seen, in the default sketch grammar, three queries cover the symmetric relation between nouns, verbs and adjectives, with the Spanish words *y|o* as coordinating conjunctions. In the Araneum Compatible Sketch Grammar, nouns, verbs, adjectives and other word categories can be matched by means of one query only. This is due to the use of regular expression notation "\!", which means negation of the attribute value, and the definition of a global condition at the end of the query, establishing that both keyword and collocate must have been attributed the same tag

[203] https://the.sketchengine.co.uk/corpus/wsdef?corpname=preloaded/pt_araneum_maius

[204] Sketch grammar for the Cetenfolha, Cetempublico corpus was also consulted. In many situations, there were no results due to insufficient data. This did not happen with the new query. Although the three sketch grammars yielded good and bad results, the option to keep this relation in the new sketch grammar does not revoke the need of further improvement in the future.

value. In my sketch grammar for Academic Portuguese, this new query captures collocates like *também e ainda* (adverbs).

It should be noted that there is a limit to how much effort can be put into finding workarounds for POS-tagging errors in my new sketch grammar definitions. Firstly, I must consider the fact that this sketch grammar is just one requirement for the development of a larger project, i.e. conceptualising and compiling the design of DOPU. Secondly, making amendments is not the solution; to definitively overcome such limitations, the tagger should be improved. But that is one important lesson to be taken from this process, namely that the quality of information provided to lexicographers, in this case through word sketches, relies not only on definitions of grammatical relations in the sketch grammar, but also on the accuracy of tools such as taggers or parsers, and also on the quality of corpus data.

7.3.4 AcadPort_v3-SkG

Since the release of AcadPortSkG in November 2016, it was edited two more times. In AcadPortSkG_v2, the previous UNARY relation `=que+verb` was split in two: `=que+verb` and `=verb+prep+que`, and a new UNARY relation was added: `=verb+conj`. In AcadPort_v3-SkG, which is the version used in the final extraction of data from CoPEP_AO90 and import into iLex, a gramrel name was modified - from `verbo se + N/suj-obj verbo se` to `%w_verbo com se + N/verbo com se+%w_N-` and tag PP3CN00 was substituted for [morphemes="se"].

Figure 7-3, Figure 7-4 and Figure 7-5 show AcadPort_v4-SkG as displayed in the Sketch Engine page. It should not be forgotten that my sketch grammar is available on the corpus tool for anyone working with Freeling-tagged corpora of Portuguese, as an alternative for the default sketch grammar currently provided.

```
#AcadPortSkG_v3 #by Tanara Zingano Kuhn Nov 2016 #Capes scholarship, process number 0973/13-0

#"V.P.*" as adjective because participle form is only tagged as verb by Freeling #edited April 2017


*STRUCTLIMIT s *DEFAULTATTR tag

*SYMMETRIC
=e_ou
1:[tag!="F.*"] [tag!="(C.*)|(F.*)"]{0,2} [lemma="e|ou"][tag!="(C.*)|(F.*)"]{0,2} 2:[tag!="F.*"] & 1.tag=2.tag

*DUAL
=Adj-Part mod %w_N/%w_Adj-Part mod N
2:"A.*"1:"N.*"
[lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")& lc!="foi|foram|fui|fomos|foste"][lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")&
lc!="foi|foram|fui|fomos|foste"]"R.*"?2:"V.P.*"1:"N.*"

*DUAL
=%w_N mod por Adj-Part/N mod por %w_Adj-Part
1:"N.*""R.*"{0,2}"A.*|V.P.*"{0,2}2:"A.*|V.P.*"

*DUAL
=%w_N ser-estar N/N ser-estar %w_N
1:"N.*" "R.*"{0,2} [lemma="ser|estar"] "A.*|V.P.*|D.*|Z.*|P.*"{0,4} 2:"N.*"

*DUAL
=%w_N ser-estar Adj/N ser-estar %w_Adj
1:"N.*" "R.*"{0,2} [lemma="ser|estar"] "R.*"{0,2} 2:"A.*"

*DUAL
=%w_N suj de V/N suj de %w_V
1:"NC.P.*"[word!="\."]{0,7}2:"VM.*3P." 1:"NC.S.*"[word!="\."]{0,7}2:"VM.*3S."

*DUAL
=%w_Adv mod Adj-Part/Adv mod %w_Adj-Part
1:"R.*"2:"A.*"
[lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")& lc!="foi|foram|fui|fomos|foste"][lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")&
lc!="foi|foram|fui|fomos|foste"]1:"R.*"2:"V.P.*"
```

Figure 7-3 Part I of AcadPortSkG_v3

```
*DUAL
=%w_Adj-Part mod por Adv/Adj-Part mod por %w_Adv
1:"A.*"2:"R.*"
[lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")& lc!="foi|foram|fui|fomos|foste"][lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")&
lc!="foi|foram|fui|fomos|foste"]1:"V.P.*"2:"R.*"

*DUAL
=%w_V mod por Adv/V mod por %w_Adv
1:"V.*"2:"R.*"

*DUAL
=%w_Adv mod V/Adv mod %w_V
1:"R.*"2:"V.*"

*SEPARATEPAGE sintagma preposicional *DUAL *TRINARY
=%w %(3.lemma) N/...%(3.lemma) %w
1:[tag="V.*"|tag="N.*"]"R.*"?"A.*|V.P.*"?3:"SP.*" "D.*|P.*|Z.*|A.*|V.P.*"{0,4} 2:"N.*"


*DUAL
=%w_V obj N/V obj %w_N
#tag corrected
[morphemes!="se"]1:[tag="V.*"& tag!="V.P.*"&lemma!="ser|estar"&lc!="foi|foram|fui|fomos|foste"&morphemes!="se"]"A.*|V.P.*|D.*|Z.*|P.*"{0,4}2:"N.*"

*UNARY
= +infinitivo
1:"V.*""VMN.*" 1:"A.*""VMN.*"

*SEPARATEPAGE preposição+Vinf *TRINARY
=...%w %(3.lemma) Vinf
1:[tag="V.*"|tag="N.*|A.*"] 3:"SP.*" 2:"VMN.*"
```

Figure 7-4 Part II of AcadPortSkG_v3

```
*UNARY
=%w_V+prep+que
1: "V.*""SP.*"[lemma="que"]

*UNARY
=%w_V+que
1: "V.*"[lemma="que"]

*UNARY
=%w_V+conj
1: "V.*"[tag="C.*"&word!="(que)|(e)|(ou)"]

*UNARY
= %w_V + Vger
1:"V.*""VMG.*"

*DUAL
=%w_verbo com se + N/verbo com se + %w_N
#name changed #tag corrected
1:[tag="V.*"&morphemes="se"]"A.*|V.P.*|D.*|Z.*|P.*"{0,4} 2:"N.*"
[morphemes="se"]1:"V.*""A.*|V.P.*|D.*|Z.*|P.*"{0,4} 2:"N.*"

*DUAL
=passiva impessoal/sujeito da passiva impessoal
[lemma="ter|haver"]?[(lc="foi|foram|fui|fomos|foste")|(lemma="ser")]1:"VMP00SM*" "A.*|V.P.*|D.*|Z.*|P.*"{0,4}2:"N.[M|C][S|N]000" [lemma="ter|haver"]?
[(lc="foi|foram|fui|fomos|foste")|(lemma="ser")]1:"VMP00SF*" "A.*|V.P.*|D.*|Z.*|P.*"{0,4}2:"N.[F|C][S|N]000" [lemma="ter|haver"]?[(lc="foi|foram|fui|fomos|foste")|
(lemma="ser")]1:"VMP00PM*" "A.*|V.P.*|D.*|Z.*|P.*"{0,4}2:"N.[M|C][P|N]000" [lemma="ter|haver"]?[(lc="foi|foram|fui|fomos|foste")|(lemma="ser")]1:"VMP00SF*"
"A.*|V.P.*|D.*|Z.*|P.*"{0,4}2:"N.[F|C][S|N]000"

*DUAL
=sujeito da passiva pessoal/passiva pessoal
1:"N.[M|C][S|N]000""A.*|V.P.*|R.*"{0,2} [lemma="ter|haver"]?[(lc="foi")|(lemma="ser"&tag="VM.*3S.")]2:"VMP00SM*" 1:"N.[F|C][S|N]000""A.*|V.P.*|R.*"{0,2}
[lemma="ter|haver"]?[(lc="foi")|(lemma="ser"&tag="VM.*3S.")]2:"VMP00SF*" 1:"N.[M|C][P|N]000""A.*|V.P.*|R.*"{0,2} [lemma="ter|haver"]?[(lc="foram")|(lemma="ser"&
tag="VM.*3P.")]2:"VMP00PM*" 1:"N.[F|C][P|N]000""A.*|V.P.*|R.*"{0,2} [lemma="ter|haver"]?[(lc="foram")|(lemma="ser"&tag="VM.*3P.")]2:"VMP00PF*"
```

Figure 7-5 Part III of AcadPortSkG_v3

7.4 Concluding remarks

It has been shown that a new sketch grammar had to be devised specifically for my research due to unsuitability of the other existing sketch grammars. Although some challenges came up during this tool development, I have no doubt that the final outcome considerably enhanced the quality of the word sketches and, in consequence, of the procedure of automatic extraction.

Chapter 8 Good Dictionary Examples – GDEX for Academic Portuguese

As demonstrated in Chapter 4, Good Dictionary Examples or GDEX is a tool in the Sketch Engine for automatic evaluation of the suitability of corpus sentences to be used as good dictionary examples. The system analyses the sentences and scores them according to syntactic and lexical features. Better-scored sentences indicate more suitability for good examples. They are thus automatically placed higher in the list of concordances, helping lexicographers with the laborious task of examples selection. In other words, "The aim is to separate good candidates for dictionary examples from the bad candidates" (Kosem et al., 2011, p.151).

In an automated extraction of example candidates from the corpus, which is our method in this research, it is possible to determine that the top **X** example candidates will be automatically extracted and imported into DWS. It is expected that these extracted sentences match a certain optimal standard as envisaged by the lexicographer in order to eliminate the need to go back to the corpus for examples selection when writing entries. Although this might sound like a tall order, it should be stressed that significant advancements of NLP tools (e.g., machine learning) have already been contributing significantly to reaching this goal.

Having said that, it should be mentioned that the context of GDEX configuration development for Portuguese has not reached such a level yet. In the Sketch Engine, there is a default configuration mostly covering sentence length, punctuation and exclusion of taboo words for the Portuguese language. Nevertheless, it is known that purpose-oriented GDEX configurations yield better results. Therefore, in the case of my PhD research, a decision was made to develop a new GDEX configuration specifically for CoPEP with the particular purpose of providing examples for DOPU entries.

This chapter gives an account of the development of GDEX configuration for academic Portuguese. It begins by providing some information on how GDEX works and setting the specific context of the development of GDEX configuration for Portuguese (8.1). Then, it moves on to a description of the first development of GDEX, which was still very basic and indicated that further measures needed to be taken (8.2). Section 8.3 reports on the customized preparation for second configuration development

and 8.4 describes the process of GDEX configuration writing. Section 8.5 wraps up this chapter with some conclusions that have pointed out what the next steps are for the continuation of developing the GDEX configuration for Portuguese.

8.1 Contextualization

According to Kilgarriff, et al., (2008, p. 426), a good example must be:

> • typical, exhibiting frequent and well-dispersed patterns of usage
>
> • informative, helping to elucidate the definition
>
> • intelligible to learners, avoiding gratuitously difficult lexis and structures, puzzling or distracting names, anaphoric references or other deictics which cannot be understood without access to the wider context. We call this its "readability".

Having that in mind, the authors set out to develop a function in the Sketch Engine in which the system analyses the concordances and automatically assigns scores, placing the highest-scored sentences at the top of the concordances list. To do that, a formula was created consisting of classifiers regarding lexical and syntactic features that contribute to the determination of a sentence's typicality, informativeness and readability, the so-called GDEX configuration. The calculation was done based on the values and weights given to these classifiers.

Initially, GDEX configuration was only available for English. Recently, however, tailor-made GDEX configurations for other languages have been developed as well, mostly with implementation of advanced NLP technology. For example, in the case of Slovene (Kosem et al., 2011) and Estonian (Koppel, 2017), machine learning techniques have been adopted in which human-judged sets of both good and bad examples were used as parameters for the optimization of classifiers' values and weight attribution. This method has tremendously increased the quality of their GDEX configurations.

Another example of a language for which customised GDEX configurations have been developed is Portuguese. In the Sketch Engine, a default GDEX configuration for Portuguese is available that was created for finding examples for the *Oxford Portuguese Dictionary* (2015). It is an adaptation of a language-independent GDEX configuration, aimed at dealing with web corpora; among a series of special

measures was the creation of a blacklist with offensive and taboo words (Figure 8-1), which were considered inappropriate for that dictionary. As such, this configuration is not academic-language oriented. Thus, to the best of my knowledge, the GDEX configuration that I have developed for automatic extraction of examples from CoPEP and import into dictionary entries in DWS for designing DOPU is a pioneer in academic Portuguese.

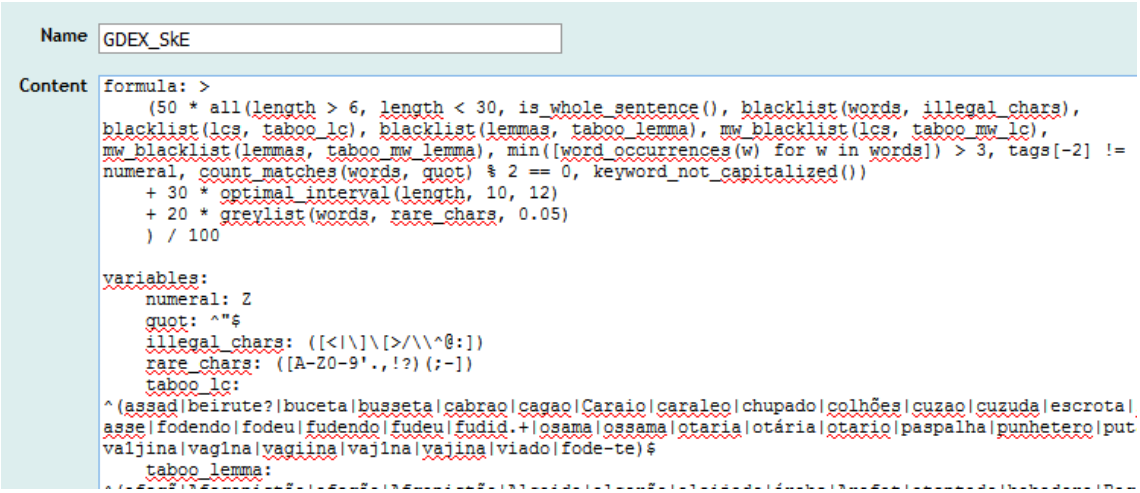


Figure 8-1 GDEX for Portuguese in the Sketch Engine

However, working on such an innovative project has its downsides. One of its major problems refers to the lack of a benchmark with parameters for comparison. If there are no clear criteria for determining what a good (or bad) example looks like, how can one evaluate candidate sentences? While other languages have created heuristics based on existing manually validated good and bad examples (as pointed out above), my development of GDEX configuration for Portuguese was still very rudimentary. It basically consisted of a series of trials with different classifiers, values and weights with evaluation of the results being made on the basis of my own subjectivity.

That was the approach I used in my first process of GDEX development for Portuguese, which will be succinctly described next.

8.2 First configuration development

The fundamental decision was to use the GDEX configuration for Slovene (version 2) as the starting point, rather than the default configuration available on the

Sketch Engine. The main reason was due to the additions offered by the former, which had introduced a number of more linguistically-refined classifiers, based on extensive studies (see Kosem et al., 2011; Kosem et al., 2013).

This first development of GDEX configuration for academic Portuguese was undertaken for the experiment of automatic extraction of data from the corpus and import into DWS. At that moment, provisional versions of tools and resources were being used (the Pre-Final Corpus and Sketch Grammar v7.1).

The starting-point was then the GDEX configuration for Slovene (version 2), comprising two kinds of penalization: hard classifiers, with Boolean conditions, i.e. if one of them is not met, penalty is very high; and soft classifiers, with weighted values for penalties and bonuses. The first test consisted of applying the same configuration as the one for Slovene, i.e. no modification in classifiers, values and weights. Hence, this is the GDEX configuration Portuguese-v1:

Portuguese-v1

- Boolean conditions (if one of them is not met, penalty is very high):

  -whole sentence

  - blacklisting spam, mainly URLs in the sentences

  - penalizing word forms (or tokens) with frequency of 2 or less

  - sentences have to be between 7-60 words long

  - keyword repetition is penalised

- Soft classifiers:

  - award to sentences between 15 and 40 (optimal length)

  - penalizing words of more than 12 characters

  - penalizing rare characters such as strange brackets etc.

  - awarding points to second collocates (10 most salient collocates of the collocation)

Alternative values and weights were tested, with new versions corresponding to the following path of modification (Table 8.1):

Table 8.1 Alternative versions of GDEX Portuguese v1

Portuguese-v2	*Portuguese-v3*	*Portuguese-v4*	*Portuguese-v5*	*Portuguese-v6*	*Portuguese-v7*
Similar to v1, but some differences: -sentences have to be between 7 and 40 words long -optimal length is 7-30 -sentences with 3 or more commas are penalised	Similar to v2, but: -sentences have to be between 7 and 30 words long -if any word occurring less than 3 times in the corpus is found in the example, the penalty is very high -penalty for any word with frequency under 50 in the corpus	Similar to v3, but: - penalty for more than 2 occurrences of "que"	Similar to v4, but: - higher weight for second collocate: 25	Same as v5, but: - penalty for any word with frequency under 100	Similar to v6, but: - penalty for any word with frequency under 200

For an analysis of the differences that each new version implemented in the example candidate sentences, a combination of word sketches and TickBox Lexicography (see Chapter 5) was the method adopted. The process consisted of performing a word sketch with concordance lines sorted by GDEX scores. Next, I would tick collocates in the word sketch results pane. Then, in a new page, examples would be provided for one chosen configuration, allowing evaluation and comparison of outcomes resulting from different configurations. Figure 8-2 is an example of such a procedure with a comparison between Portuguese-v1 and Portuguese-v7.

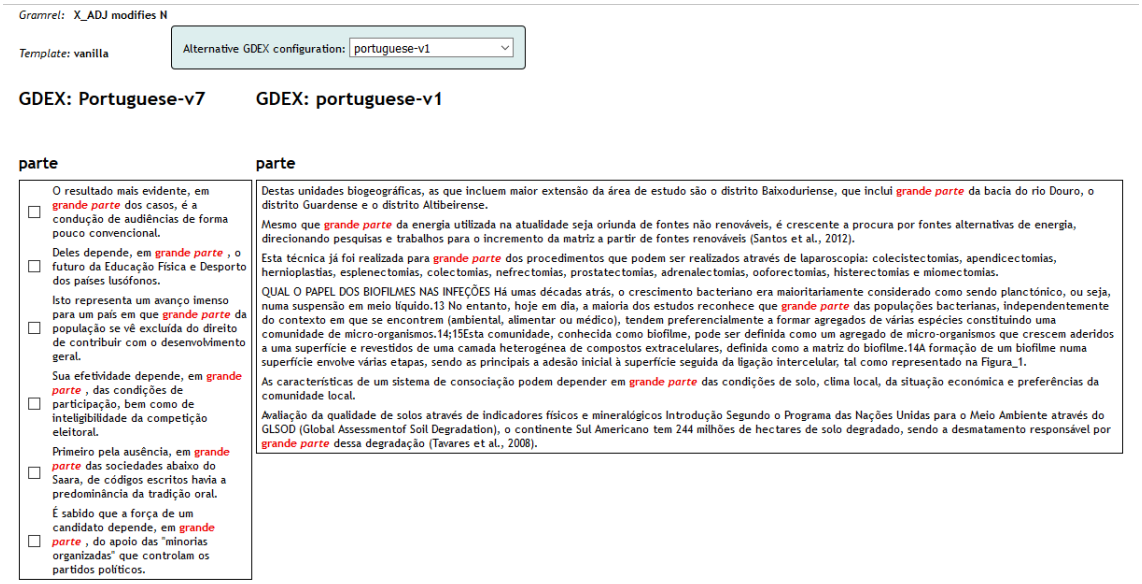


Figure 8-2 TickBox Lexicography for GDEX configuration development

This first trial with GDEX configuration development was an important initial step towards a continuous process of GDEX configuration for academic Portuguese. I learnt that when radically different values are introduced, as demonstrated by the comparison between Portuguese v1 and Portuguese v7, impacts on the results are more visible, while minor tweaking, although producing differences, are more difficult to spot and measure. In any event, to measure the extent to which these differences improve the quality of the sentences as good dictionary examples one fundamental condition must be met: to have parameters that define what a good dictionary is so that comparison can be undertaken.

As previously mentioned, there are no assessed sets of good or bad dictionary examples for Portuguese, let alone for academic Portuguese. Indeed, it has been shown that the most appropriate manner to obtain parameters for judgment of examples for DOPU target users would be verified examples extracted from CoPEP.

Nevertheless, carrying out the study for obtaining human-validated parameters for GDEX configuration development was out of the scope of my PhD research. Another alternative would be to use metrics related to readability, like the Coh-Metrix-Port (Scarton & Aluísio, 2010) for BP and the LX-CEFR[205] (Branco, Rodrigues, Costa, Silva, & Vaz, 2014) for EP. The former has been used, among other things, for text

[205] http://nlx.di.fc.ul.pt/~jrodrigues/camoes/caracteristicas.html

simplification and summarization, while the second aims to support automatic classification of texts, thus helping instructors to make principled decisions for material selection for proficiency exams.

Given the different purposes of those metrics and their reference to texts, not sentences, application of the measurement results should be dealt with carefully. But it should be interesting to experiment to measure CoPEP according to those metrics, provided that they can be compared to metrics from attested good and bad examples – here, again, the point being the lack of a benchmark. Future developments should include this method of corpus analysis.

For the project, a provisional solution was found in which statistics on average sentence length, word length, among others, are obtained from CoPEP and used as references for setting up parameters in my GDEX configuration. This will be shown in next section.

8.3 Preparation phase

In the second attempt to devise GDEX configurations for academic Portuguese, the point of departure was already a modified version of GDEX for Slovene (version 2), namely, Portuguese v7, with definition of values for classifiers based on statistical calculations from CoPEP. Moreover, instead of using word sketches and TickBox Lexicography for comparison and evaluation of configurations, the procedure of GDEX configuration development is now conducted in the GDEX editor[206] in the Sketch Engine (currently provided as a beta version). This tool requires the usual configuration file comprising classifiers and values, as well as sample word sketch corpus query language (CQL) searches.

8.3.1 CoPEP statistics calculations

Due to the lack of statistics of manually validated good and bad examples to serve as parameters, a decision was made to utilize statistics on CoPEP. Although not ideal, it was a workable alternative and provided a sound point of reference to support further manual configuration development.

[206] https://beta.sketchengine.co.uk/gdex_editor

Drawing on language features measured for other GDEXes, I established what could be reused and what is more specific to CoPEP and DOPU design purposes. The metrics below were then calculated for each great area of knowledge in each language variety subcorpus. The codes between brackets correspond to the part-of-speech tag in the Freeling tagset, which were required for the statistics script.[207]

- Average sentence length
- Number of sentences with one word, two words, etc.
- Average word length
- Number of words with one letter, two letters
- Average of longest words in the sentences
- Average of shortest words in the sentences
- Frequency ranges of words in sentences according to the frequency list[208]
- Type/token ratio
- Number of sentences starting with capital letter and finishing with full stop (Fp)
- Number of sentences starting with capital letter and finishing with exclamation mark (Fat)
- Number of sentences starting with capital letter and finishing with question mark (Fit)
- Number of relative pronouns (PR.*) in a sentence
- Number of "*se*" (PP3CN0) in a sentence
- Number of demonstrative pronouns (PD.*) in a sentence
- Number of indefinite pronouns (PI. *) in a sentence
- Number of VM. *1P. (verb first person plural) in a sentence:
- Number of V.*(any verb) in a sentence:
- Number of verbs+clitics [tag="V.*" & word=".*-.*"] in a sentence[209]
- Number of adverbs (R.*) in a sentence
- Number of common nouns (NC. *) in a sentence
- Number of proper nouns (NP. *) in a sentence

[207] This script was prepared by a hired computational linguist.

[208] A lempos frequency list was provided.

[209] Calculation of this metric returned empty results.

- Number of adjectives (A.*) in a sentence:
- Number of commas (Fc) within a sentence:
- Number of quotation marks (Fe) within a sentence:
- Number of semicolons (Fx) within a sentence:
- Number of colons (Fd) within a sentence:
- Number of open quotation open/close quotation (Frc/Fra) within a sentence:
- Number of articles (DA.*) in a sentence:
- Number of indefinite determiners (DI. *) in a sentence
- Number of demonstrative determiners (DD.*) in a sentence
- Number of coordinating conjunctions (CC) in a sentence
- Number of subordinating conjunctions (CS) in a sentence
- Number of prepositions (SP) in a sentence
- Number of numbers (Z) in a sentence

Table 8.2 shows part of the results. In Appendix D (see CD-ROM), the full table is available. Some of the interesting findings revealed by these statistics will be discussed in Chapter 12.

Table 8.2 Partial results of statistics on CoPEP

	A	B	C	D	J	N	P	AS
1	great_area	variety	Sentences	Sentence length	Sentence question	Word max	adjective	Frequency [101,140]
2	Agricultural Sciences	Eu	34908	37,105	0,6%	13,024	2,896	0,226
3	Health Sciences	Eu	196282	34,359	0,9%	13,105	3,015	0,176
4	Applied Social Sciences	Eu	74142	36,358	1,5%	12,875	2,94	0,16
5	Human Sciences	Eu	349224	36,75	2,0%	12,897	3,134	0,183
6	Exact-Earth Sciences	Eu	11481	35,762	0,3%	11,669	2,691	0,221
7	Engineering	Eu	10834	30,838	0,5%	12,145	2,105	0,172
8	Agricultural Sciences	Br	37840	34,786	0,0%	12,889	2,424	0,223
9	Health Sciences	Br	193591	35,114	1,4%	12,986	2,575	0,141
10	Applied Social Sciences	Br	77374	34,864	0,9%	13,208	2,675	0,132
11	Human Sciences	Br	377169	33,843	2,0%	12,533	2,935	0,175
12	Exact-Earth Sciences	Br	13004	30,776	0,4%	13,486	2,294	0,249
13	Engineering	Br	11216	31,27	0,3%	12,859	2,129	0,18

8.3.2 Word Sketch Corpus Query Language Searches

As shown above, initially, the method for GDEX configurations development in the Sketch Engine included a combination of word sketch results and Tickbox Lexicography (see Chapter 5). The process consisted of performing a word sketch with concordance lines sorted by GDEX scores. Next, lexicographers would tick off collocates in the word sketch results pane. Then, in a new page, examples would be provided for one chosen configuration, allowing for evaluation of outcomes by comparison with an alternative configuration, as shown in Figure 8-2 TickBox Lexicography for GDEX configuration development.

Since 2014, however, GDEX configurations can be developed in a much more user-friendly interface, which is available at the moment in a beta version. As mentioned earlier, for this new procedure, in addition to a configuration file comprising classifiers and values, sample word sketch corpus query language (CQL) searches were also required; these searches are structured as [ws(“keyword”, “relation”, “collocate”)], with keyword having the form of lempos, i.e. lemma with POS-tag indication, as in *estudo*-n (‘study’; -n stands for noun).

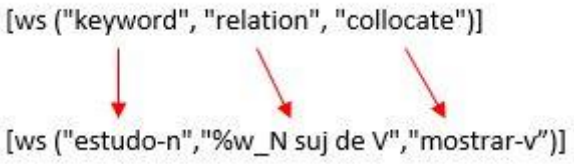
Writing word sketch CQL searches consisted of the following steps, for different lemmas of each word category, namely, noun, verb, adverb, and adjective.

a. Choose one high frequent lemma (lempos) as keyword;
b. Do a word sketch;
c. Make a record of the top collocate from each relation;
d. Substitute elements of a word sketch CQL query with the lempos, relation and collocate in question.

Let us use the noun *estudo* ('study') as an example. A partial word sketch result is seen in Figure 8. For the grammatical relation *estudo_N suj de V* (noun subject of verb), the top collocate is the verb *mostrar* ('to show').

Figure 8.2 Partial word sketch result of *estudo*

Substitution was as follows:



For the 44 grammatical relations in word sketches, 43 sample lemmas were used: 15 verbs, 13 nouns, eight adverbs, and seven adjectives (see Appendix E).

8.4 Writing GDEX configuration for CoPEP

GDEX configuration development for academic Portuguese began with basic features, resulting in a provisional set of criteria (Portuguese-v7) that was used for the

experiment of automatic extraction (Chapter 9).[210] It consisted of the same classifiers as the GDEX for Slovene (version 2), with the addition of a new classifier (penalization of more than two occurrences of *que*) and adjustment of values and weights. The current version - AcadPort-4_GDEX -configuration of GDEX resulted from the procedure described below, in which statistics from CoPEP were used as a reference and the new GDEX editor was used.

GDEX configuration development for CoPEP adopted a rule-based approach as machine-learning techniques could not be used due to lack of manually-validated data sets of good (and bad) examples.

In the GDEX editor interface (Figure 8-3), two alternative GDEX configurations can be seen side-by-side, enabling quick classifiers tweaking, whose influence can be immediately measured in the example[211] outcomes located in the lower part of the panel, beneath the settings. These examples are ranked and displayed next to each other, with GDEX scores, thus allowing simple visualization and greatly facilitating comparison.

[210] As mentioned earlier, this experiment was performed on the Pre-final corpus, which was an unfinished version of CoPEP (not fully balanced nor post-processed). Two of the prerequisites for the procedure, Sketch Grammar and GDEX configuration, were thus initially developed on this corpus version. For the final automatic extraction of data and import into DWS, as described in section 9.2, the latest versions of all the required resources were used.

[211] When developing GDEX configurations, we are dealing with (corpus) sentences and example candidates; good examples are those selected and saved in the database. However, as a matter of simplification, in this chapter, sentences and examples are used interchangeably and refer to potential examples for the dictionary.

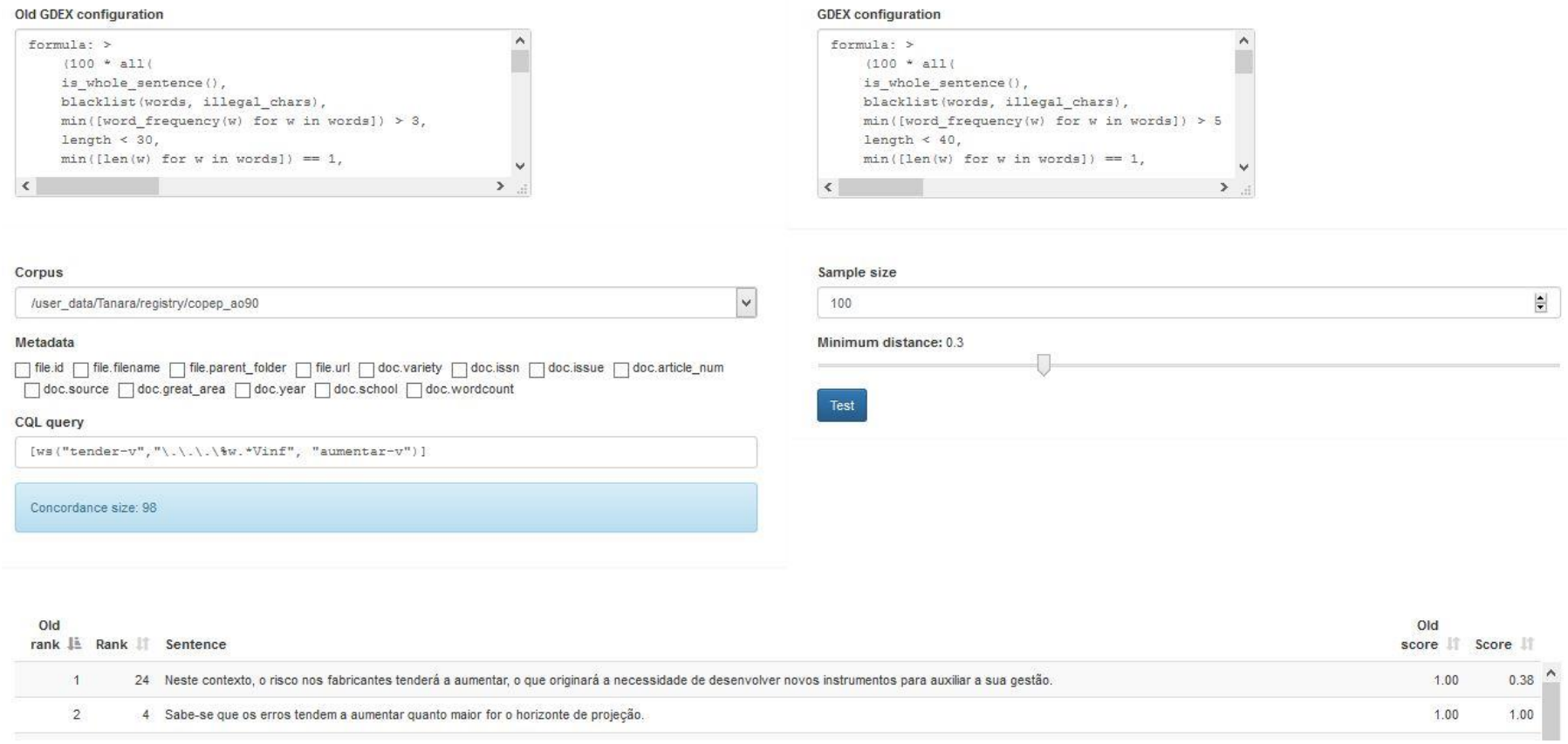


Figure 8-3 GDEX editor interface

As can be seen, at the top there are two fields, one for each GDEX configuration. Users simply paste the configuration and can edit values in the field. A drop-down box shows available corpora; once a corpus is selected, its metadata is displayed with tick boxes for additional information to appear next to the sentences. The word sketch CQL queries (as shown in 8.2.2 above) are pasted in the CQL query field. Sample size concerns how many sentences will be presented, while the Minimum distance sliding bar is for defining the difference between the examples, measured in the Jaccard similarity index, set at 0.3% by default (Kosem et al., forthcoming).

In the results, rank values are presented to the left of the sentences, whereas scores are seen to the right; both can be used for lines sorting. With this information, detailed comparisons of the influence of changes in each classifier, value, and weight on the quality of examples can be performed, and adjustments can be made accordingly.

The procedure of writing a GDEX configuration for academic Portuguese thus involved a series of iterations with experimentation of varying values and weights.

The latest configuration version is AcadPort-4_GDEX, as shown in Figure 8-4.

```
formula: >
    (100 * all(
    is_whole_sentence(),
    blacklist(words, illegal_chars),
    min([word_frequency(w) for w in words]) > 50,
    length < 40,
    min([len(w) for w in words]) == 1,
    blacklist(words, spam),
    keyword_repetition(lemmas) == 1,
    length > 7
    )
    + 10 * legacy_optimal_interval(length, 10, 30)
    + 1 * greylist(words, longer_than_twelve, 0.1)
    + 1 * greylist(words, rare_chars, 0.1)
    + 1 * greylist(words, mixed_symbols, 0.1)
    + 5 * max(0, 1 - sum([0.2 for lemma in lemmas if lemma_frequency(lemma) < 500]))
    + 5 * (1 - 0.2 * max(0, count_matches(words, comma) - 2))
    + 9 * (1 - 0.2 * max(0, count_matches(words, que) - 2))
    + 30 * min(1, sum([0.5 for score in lemma_collocation_scores(fromw=-5, tow=5,
    minfreq=5, mincnt=3, maxitems=10,
    colfunc='PROD_MI.LOG_F')[max(0,kw_start-5):kw_end+5] if score > 0]))
    ) / 162

variables:
    illegal_chars: ([<|\][>/\\])
    spam: (..*@..*\...*|http:/|.\..*\..|Abstract|Acknowledgments|Creative_Commons_License)
    longer_than_twelve: (............)
    rare_chars: ([\[\]"'\\\*»«/ _–#“”•@\|=’~$×}{)(;:-])|(\b\.\.\.\b)
    mixed_symbols: ([-a-zçA-ZÇ][0-9.,/;\\!@$%^*\[\]()_ ]|[.,/;\\!@$%^*\[\]()_
    ][a-zçA-ZÇ]|[-,/;\\!@$%^*()\[\]_ ][-.,/;\\!@$%^*()\[\]_ ]|[\[\]<>^])
    que: (que)
    comma: (,)
```

Figure 8-4 AcadPort-4_GDEX configuration

Classifiers (including values) from the GDEX for Slovene (version 2) that had been kept unchanged in Portuguese v7 were: a) hard penalization: whole sentence, illegal characters, minimum word length (1), minimum sentence length (>7), keyword repetition; b) soft penalization: words longer than 12 characters, rare characters, and sentences with more than two commas. While most of these classifiers are language-independent, common values for Slovene and Portuguese were found with regard to minimum sentence length, given that in CoPEP, too, shorter sentences tend to lack context; number of characters per word, since average word length in the CoPEP corpus is 12.81; and penalty for more than two commas in a sentence due to syntactic complexity indicated by use of several commas.

The novelties implemented in this new GDEX configuration development refer to the following issues. Firstly, the analysis of the CoPEP corpus showed that average sentence length is 34.32 tokens, thus maximum sentence length was set to 40 tokens. Secondly, the optimal interval of the sentence was set to between 10 and 30 tokens, hence examples within this range received bonus points. Thirdly, spam variables were adapted to address some noise still found in CoPEP, e.g. Creative_Commons_License. Fourthly, mixed symbols were adjusted, with typical Slovene letters (e.g. č) being substituted for the letters from the Portuguese alphabet that are not found in the English alphabet.

Similarly to penalization of occurrence of more than two commas in an example, the use of more than two words *que* (‘that/which’) was penalised in Portuguese v7 and kept in this new version. This is because when used as a relative pronoun or subordinate conjunction, *que* contributes to higher syntactic complexity (and usually length) of the sentence. This is a characteristic that should be avoided in examples, as it might hamper understanding.

Classifiers with the greatest influence on the scoring were minimum frequency of any token in a sentence, minimum frequency of lemmas, and second collocate. The first two classifiers were given higher values in comparison to the Slovene configuration, with minimum frequency of any token in a sentence set to 50, and minimum frequency of lemmas set to 500. With this combination, less frequent acronyms, for instance, were avoided in examples, making them more suitable for less

course-experienced students. Bonus points awards for second collocates was adjusted, both with a higher weight for this classifier and higher bonus points for each collocate.

It was decided that classifiers for keyword position and sentence tag initials (as in the GDEX configuration for Estonian) would not be included in AcadPort-GDEX due to lack of parameters against which to proceed with evaluation.

8.5 Concluding remarks

Despite the limitation posed by the lack of previously human-validated analyses of examples to provide workable statistics, it has been shown that the tentative alternative, namely, the use of the statistical description of DOPU, was undoubtedly helpful.

The next steps in the development of a GDEX configuration for academic Portuguese involve using an upgraded function in the GDEX editor where the impact of each classifier can be measured, thus facilitating the evaluation of the results. It should be possible to verify whether different word classes require different classifiers/values, as is the case with Slovene (Kosem et al., 2013).

More importantly, before implementing these new functions, it is key to carry out a study where good and bad examples are identified and can be used as a benchmark for GDEX configuration development. One of the latest strategies for such a type of research is crowdsourcing (see Fišer & Čibej, 2017). One possible avenue is to extract a number of examples based on the current GDEX configuration and to ask people (ideally, the future target user of DOPU) to read sentences and evaluate their quality by confirming if the meaning, or use, or syntactic behaviour (we should decide on that later) is clarified with that example. The results would provide parameters for further configuration tweaking.

Part III

THE DICTIONARY DESIGN PROPOSAL

Introduction to Part III

In part I, I presented a call for an online corpus-driven dictionary of Portuguese for university students. I argued for the socio-political relevance of such a tool, particularly in the context of Brazil and Portugal, but also in reaching out towards the broader scope of CPLP. Moreover, I showed that academic Portuguese is a register of language with its own characteristics thus requiring specific research-informed teaching and learning materials; however, a dictionary for these purposes have not been created up to date to my knowledge. I then defended that the rapid increase in lexicography-related technology has made a lexicographical project such as DOPU a perfectly feasible enterprise. I finished the first part with the presentation of a principled plan for developing DOPU.

Part II described the process of setting up the semi-automated approach to dictionary-making, which mostly refers to assembling what is needed for application of the method of automatic extraction of data from the corpus and import into DWS. As this was the first time that such an approach had been employed in the making of a dictionary of Portuguese, I had to develop specially-tailored tools and resources for the design of DOPU. I then compiled a carefully-designed corpus of academic texts – the CoPEP; devised a new sketch grammar for CoPEP; and developed new GDEX configurations based on the characteristics of my corpus and the purpose of the examples.

Now, in Part III, I propose a design of DOPU. For that, I passed through the stages of acquiring lexicographic evidence, building candidate headword lists, and compiling entries, guided by the plan presented in Chapter 4, and applying the resources and tools developed in Part II. Thus, in Chapter 9, I report on the procedure of automatic extraction of data from CoPEP and import into iLex. Given that the evaluation of the procedure required actual entry compilation, attention is drawn to the fundamental change in the lexicographical work addressed by the semi-automated approach. Now, the point of departure for entry compilation is directly the DWS, which had been automatically populated with pre-defined data from CoPEP. I will demonstrate how this new way of writing entries works with some illustrative cases. Then in Chapter 10, I cover matters related to the macrostructure of DOPU, exploring the guidelines laid

down in Chapter 4, this time bringing in actual data to support some decision-making. Finally, in Chapter 11, I refer to the microstructure of DOPU, with the presentation of the elements and some illustrations with actual data from CoPEP.

Chapter 9 Automatic data extraction

In this thesis, automatic extraction of data refers to the extraction of predefined types of information from CoPEP in the Sketch Engine via a Python API script. The following data were automatically extracted:

- Lemmas
- Word class
- Frequency
- Grammatical relations
- Collocates
- Examples

These data were then imported into the iLex DWS, automatically populating the entries' elements according to the types of information.

The extraction procedure conducted in this PhD research used the Slovene API script as a point of departure, introducing important project-oriented modifications, as will be shown below. Implementation of the process closely followed the steps described in Kosem et al. (2013) and Gantar et al. (2016).

This chapter reports on the application of the procedure. First, an experiment with a set of sample lemmas was carried out, utilizing pre-final versions of the required especially-built resources and tools in order to evaluate this method as a means to provide lexical content that would serve as a basis for compiling entries of DOPU. Evaluation of the outcomes indicated that the method could be used for the purpose of developing DOPU. However, improvement of the resources was considered necessary to better cater for factors such as spelling variance in the corpus. The experiment is described and evaluated in section 9.1 below.

A second extraction was then performed on enhanced versions of the resources, utilizing a very similar set of sample lemmas. The main objective was to verify whether the improvements introduced to reduce problems stemming from spelling variation in the corpus were effective. Comparison between outcomes from each extraction confirmed the validity of the employment of more qualified resources and tools. Moreover, it pointed out that procedure upgrades could potentially contribute to further

facilitate the lexicographer's work. Description of the second extraction procedure and outcomes evaluation are presented in 9.2.

9.1 The experiment

A modified version of the Slovene API was used for extraction of the data from CoPEP.[212] More specifically, my procedure introduced two important additions. The first one was the inclusion of additional information provided by the clustering and longest-commonest match (Kilgarriff et al., 2015) functions in the Sketch Engine; this information was added to the data after the extraction, at a post-processing stage. The main aim was to assist lexicographers in grouping collocates and in identifying multi-word expressions, as well as facilitating the detection of incorrect information.

While the first addition was language non-specific, i.e. it can be used in automatic extraction for other languages, the second was specific to Portuguese. The coexistence in CoPEP of two language varieties, i.e. BP and EP, following three different spelling conventions, i.e. FO43 (Brazil before 2009), AO45 (Portugal before 2009), AO90 (Brazil and Portugal after 2009) (see Chapter 6 for an illustrated explanation), together with the fact that DOPU aims to equally represent BP and EP, posed challenges for data extraction. These were aggravated as attempts were being made to assign variety labels not only to headwords, but also to collocations, and if relevant, to grammatical relations. A decision was made to extract data from both variety subcorpora separately, and also add statistics on grammatical relations and collocations from the whole corpus.

The objective of this experiment was to evaluate the validity of this method for the compilation of entries of DOPU. The extraction was conducted on an earlier version of CoPEP, the Port-Acad-pre-final corpus (henceforth the Pre-final corpus), with almost the same size (45 million words). The main difference was that in the Pre-final corpus, language variety balancing had not been carried out (see Chapter 6, section 6.2.2.2.3.3). This means that the criterion temporarily adopted for building BP and EP subcorpora

[212] This experiment was carried out during a Short-Term Scientific Mission at the University of Ljubljana, under the supervision of Dr. Iztok Kosem. I was given access to an individual licence to iLEX , to which I am very grateful. For this reason, the iLex layout is in Slovene.

was the source of the texts, i.e. SciELO Brazil – Brazilian Portuguese; SciELO Portugal – European Portuguese (see details in Chapter 6).[213]

The experiment involved three phases: 1. preparation for the procedure; 2. process of extraction of data and import into iLEX; and 3. evaluation. Each phase is described in the next subsections.

9.1.1 Preparation phase

This phase involved the provision of requisites for the procedure of automatic data extraction as determined in Kosem et al. (2013, pp. 35-36) and Gantar et al. (2016, pp.213-217) (9.1.1.1 below), and the creation of solutions that automatically handle spelling variance introduced by the coexistence of language varieties before data is imported into DWS (9.1.1.2).

9.1.1.1 Parameters settings

a) A selection of lemmas for extraction

My aim with the selection of lemmas for extraction was to choose a sample that would expose as much and varied information as possible. Thus, the criteria adopted were frequency and heterogeneity of lemmas, with special attention to those words that could cause problems, such as language variety variants.

In the Slovene extraction experiment, the definition of maximum collocation frequency for lemmas was required in order to limit data to a manageable size, given that extraction was conducted on the 1.2 billion-word Gigafida corpus. By contrast, in my experiment, the concern referred to the extraction of enough data due to the small size of the corpus.

For instance, it was noticed that low-frequency lemmas are ineffective for evaluation of the outcome due to two reasons. Firstly, they yielded few word sketches, providing little information for output analysis. Secondly, the frequency of structures and collocates in many cases were under the cut-off threshold that had been set (6 for both, as will be shown in item *e* below), meaning that they would not be extracted.

[213] When planning the construction of CoPEP, a definition was made to run automatic identification of language variety (with further manual review of critical cases) on top of text sources as it is widely known that correspondence between place of publication and language variety of a text is not necessarily one-to-one in the case of the Portuguese language.

Figure 9-1 illustrates a situation where only one relation would have been extracted had this lemma been included in the sample lemma list.

Figure 9-1 Partial word sketch result of *consertar* (verb)

As can be seen, *consertar* (v) ('to fix') occurs 40 times in the corpus, 17 of them in the verb+object relation. The remaining occurrences are scattered among different relations that are below the cut-off point (set at 6, as will be shown below). For evaluation purposes, the only structure and collocate meeting the minimum frequency parameters for extraction are verb+object and the collocate *máquina* ('machine'), as can be seen in the word sketch result (Figure 9-1). The conclusion is that having only six occurrences, under only one grammatical relation, is not productive.

Thus, a decision was made to select high and mid-frequency lemmas, which provided "good-sized word sketches" (Kosem et al., 2013, p. 36), that is, word sketches offering a variety of collocates and examples.

The second criterion for lemma inclusion in the sample list was diversity of characteristics. For that, lemmas belonging to four word categories, namely, noun (85), adjective (85), adverb (45) and verb (85), and with spelling variations were selected. For the latter, three different kinds of variation[214] were included, as shown in Table 9.1:

[214] See Chapter 6 for a full account on orthographic variation and spelling reform.

Table 9.1 Kinds of orthographic variations selected for the experiment and some examples

Kind of orthographic variation	Examples
Between BP and EP; extinguished with AO90 (accent-related)	('frequently', adverb) *freqüentemente* - Brazil, FO43 *frequentemente* - Brazil, AO90; Portugal, AO45 and AO90
Between BP and EP; resulting from the reform	('aspect', noun) *aspecto* - Brazil, FO43 and AO90; Portugal AO45 *aspeto* - Portugal, AO90
Between BP and EP; not changed by the reform	('economic', adjective) *econômico* – BP *económico* - EP

Lemmas should be recorded as lempos in the list for extraction. As previously explained, lempos are lemmas with POS-tag, for instance, *estudo-n*, in which the first part is the equivalent to the noun 'study' in Portuguese and the hyphenated tag indicates the word class of that lemma.

b) Finely-grained sketch grammar, designed specifically for the purposes of automatic extraction

For the experiment, sketch grammar version 7.1 was used.[215] It consisted of *UNARY, *DUAL, *TRINARY and *SEPARATEPAGE directives. As UNARY relations were not extracted, the list of relations provided for the procedure comprised 8 gramrels for adjectives, 5 gramrel for adverbs, 10 gramrels for nouns and 27 gramrels for verbs, in addition to all possible relations yielded by *SEPARATE PAGE, which consisted of prepositional phrases as complement of noun, verb or adjective as keywords.

c) GDEX configurations.

[215] Chapter 7 reported on the process of sketch grammar development for academic Portuguese, which included a number of modifications, from version 1 to the current AcadPortSkG_v3.

At this time, I was using GDEX Portuguese v7.[216] That was the result of the first round of experimentation with the GDEX configuration for Slovene (version 2), which included some tweaking of soft and hard classifiers.

d) Preparing a Python API script

The API script for extraction of data from the Pre-final corpus was built on the Slovene script and prepared by a contracted programmer. A slight modification of the procedure originally set up for Slovene was required due to the inclusion of LCM information, collocate clustering and assignment of variety label. This information was added to the extracted data at a postprocessing stage, as will be shown in more detail in the next subsection.

e) Setting parameter values for the API script

A major asset of this method is that the computer is responsible for the selection of the relevant data to be extracted from the corpus, facilitating lexicographers' work by leaving out, for instance, non-typical collocations and bad examples. Nonetheless, for the system to know what to include, parameter values have to be set by the lexicographers working on the project.

As already mentioned, one of the greatest challenges of the application of innovative techniques to develop my PhD project is the lack of a previous reference for the Portuguese language. This is also the case with parameter setting. Thus, it was decided that the following values were a good starting point. Ideally, this experiment will be continued and tests will be done to fine-tune values and resources.

The parameter values were:

- Minimum frequency of collocate: 6
- Number of examples per collocate: 3
- Minimum frequency of grammatical relation: 6
- Minimum salience of a relation or collocate: 0

With the exception of LCM, which is automatically provided by the Sketch Engine, the other new additions to the procedure (mentioned earlier) required setting

[216] GDEX configurations development were also described earlier in Chapter 8.

extra parameter values. Since extraction of this kind of data has never been attempted, not even for Slovene, the solutions put forward were highly experimental and, undoubtedly, further investigation will contribute greatly following the publication of this thesis.

For clustering values, a test was performed with some lemmas. The default value in Sketch Engine is 0.15 of granularity. Clustering collocates were tried out with different values and I analysed the results, verifying semantic proximity of clustered collocates.

One example is shown in Figure 9-2. The lemma (keyword) under scrutiny was the verb *fazer* (make/do) and the grammatical relation was *noun subject of verb*. The word sketches show the collocates found, that is, the nouns that are subjects of the verb *fazer*. With a cluster value of 0.25, the clustering around the collocate *paciente* included *doente* and *adolescente*. *Paciente* and *doente* are variants of the same lexeme ('patient', noun), with the first being used in Brazil and the second in Portugal. Thus, *adolescente* ('teenager') only shares the trace 'human' with the two other nouns, indicating that clustering value was not ideal yet. When increasing granularity to 0.30, the *paciente* clustering no longer included *adolescente*, which was grouped under a one-item cluster. Evidently, not all collocates always share the same semantic traces, as clustering is calculated like the Sketch Diff function (see Chapter 5), that is, based on statistical calculations, not semantic values.

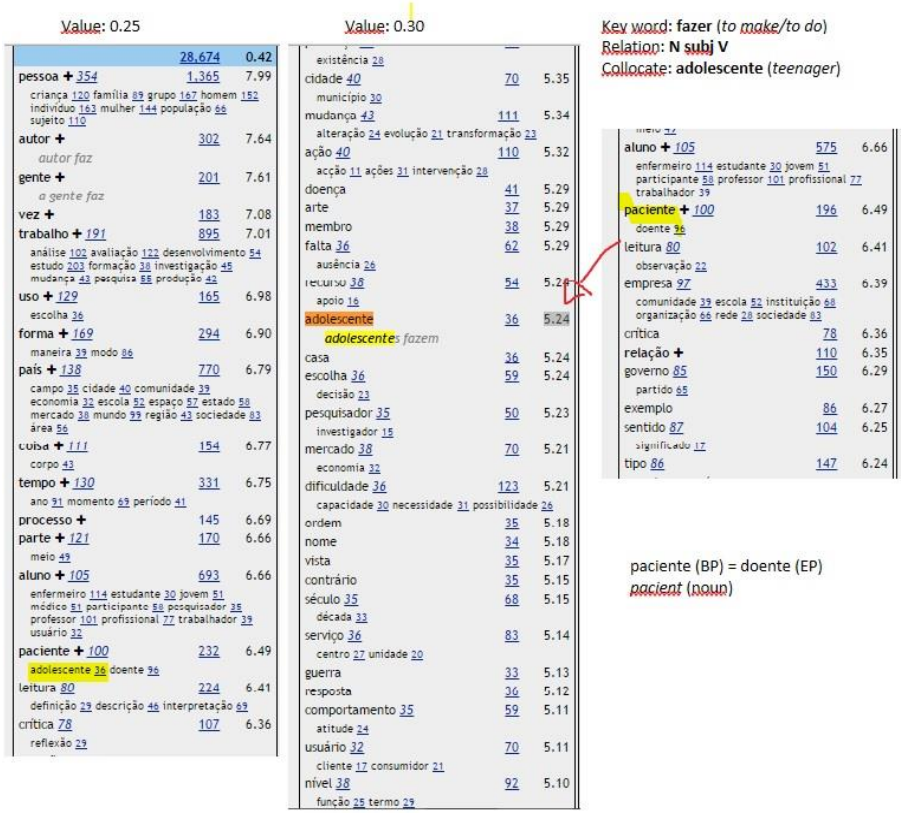


Figure 9-2 Cluster values testing

9.1.1.2 Dealing with coexistence of language varieties in one corpus

In the case of the Portuguese language, treatment of language varieties in multi-varietal corpora has been usually overlooked (see Chapter 6). Moreover, the lexicographic tradition in Brazil and Portugal is to create dictionaries of Portuguese concerning their national varieties; in fact, more often than not, this issue is not even addressed. Given this situation, I do not have alternative experiences to resort to in order to verify how coexistence of language varieties in one corpus or one dictionary has been tackled.

One dictionary that explicitly states that it covers both BP and EP varieties is the *Oxford Portuguese Dictionary* (2015). This is a bilingual corpus-based Portuguese/English- English/Portuguese dictionary which adopted the Sketch Engine as a corpus tool, also using functions such as word sketch and GDEX for lexical analysis.

As our projects have significant points in common, it seemed sensible to verify how this issue was handled by OUP. According to Kilgarriff, Pomikalek, Jakubíček, & Whitelock (2012), the Keyword function (see Chapter 5) was used to create a list with

the distinctively BP or EP lemmas found in the corpus, thus helping lexicographers with labelling. Another solution introduced in this lexicographical project was the possibility of providing information about variety typicality already in the word sketch. Called 'hypothesis', the mechanism involved statistical calculations that answer to "Is the word one of the top *x* % most-Brazilian words? Is the word one of the top *x* % most-European words?", with *x* said to be probably 0.5[217] (Kilgarriff et al., 2012, p. 784).

Unlike the web crawled corpus employed in the OUP project, in which the language variety of each text could not be firmly attested (there was a reference to URLs only), the Pre-final corpus contained texts from known sources (Brazilian and Portuguese academic journals) which were annotated as metadata in each document.[218] This particularity of my corpus enabled the creation of an alternative solution in which calculations of frequency statistics were used for the assignment of variety labels at headword, grammatical relation (also called structure), and collocate levels. For that, it was necessary that data from each subcorpora were extracted separately.

The mathematics were computed in the following way:

- if a headword/structure/collocate has frequency 0 in one subcorpus, set @variety to the other subcorpus;
- taking headword/structure/collocate frequency in each subcorpus, if frequency ratio between the subcorpora is greater than 15.0, set @variety to "typical" followed by the subcorpus name.

This information was added to the extracted data at a postprocessing stage. The figures below present examples of how that information is displayed to the lexicographer in the iLex DWS[219], both using XML visualization and entry view visualization. Figure 9-3 gives an example of variety label assignment, while Figure 9-4 shows a typical variety label assignment. The keyword is *académico*, which is the EP spelling of 'academic'.

[217] It is not known, however, if that was the value that ended up being adopted.

[218] At that point of the research, a correspondence was made between source and variety. Nevertheless, it is known that this is not a solid criterion, so the final corpus version, the CoPEP (see Chapter 6), undertook a language variety identification process that led to more accurate text annotations.

[219] This experiment was carried out during a STSM at the University of Ljubljana, under the supervision of Dr. Iztok Kosem. I was given access to an individual licence to iLEX , to which I am very grateful.

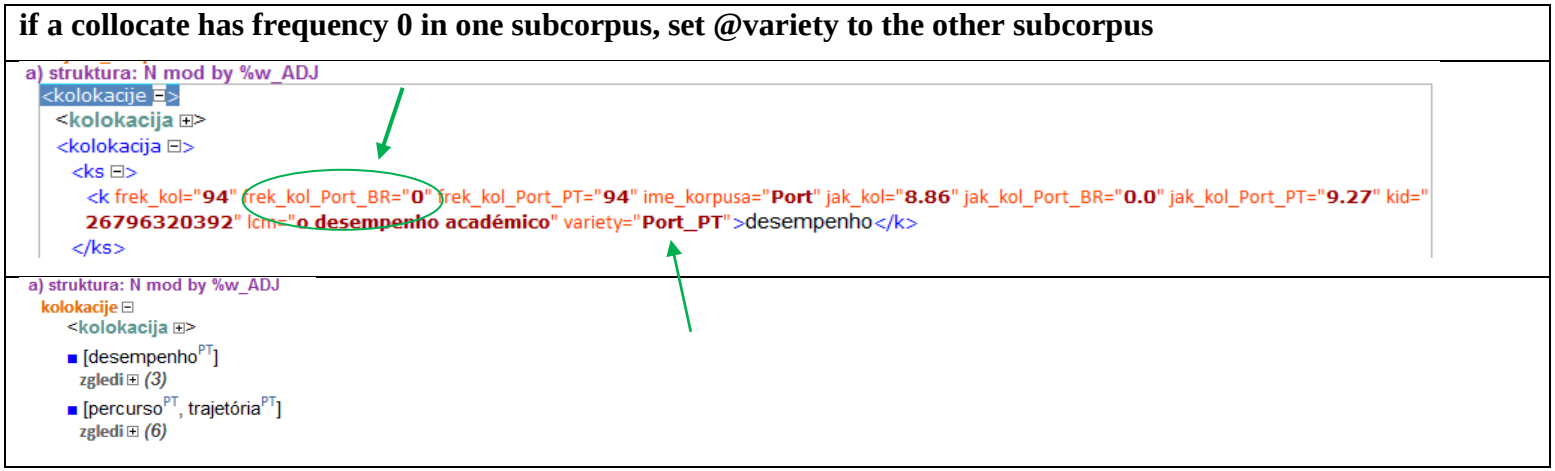


Figure 9-3 Variety label assignment

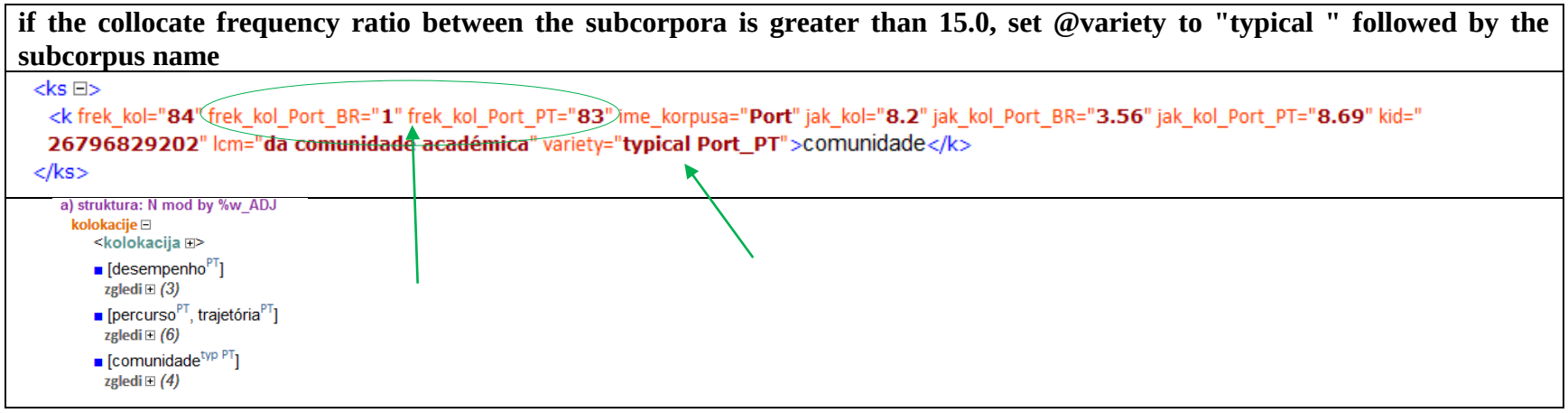


Figure 9-4 Typical variety label assignment

9.1.2 Extraction process

As presented above, these were the prerequisite and parameter values defined for the procedure:

- Pre-final corpus
- 300 lemmas
- Sketch Grammar version 7.1
- GDEX configuration Portuguese v7
- Minimum frequency of collocate: 6
- Number of examples per collocate: 3
- Minimum frequency of grammatical relation: 6
- Minimum salience of a relation or collocate: 0
- Clustering value: 0.30

The extraction procedure (Table 9.2[220]) consisted of extracting data from each subcorpus and the whole corpus, merging datasets, adding clustering and LCM information, and then importing the data into the iLex DWS.

Table 9.2 Procedure of automatic extraction

STEP	PROCEDURE AND OUTPUT
1. Extraction of information from each variety subcorpus	Extraction of relations, collocates and examples for sample set of lemmas
	Extraction of all collocates per relation from each subcorpus (without examples) to obtain frequency and salience info
2. Merging datasets	Merging all sets of data
3. Adding clustering info	Extracting clustering data and clustering extracted collocates from step 2
4. Adding longest-commonest match (LCM) info	Extracting LCM information and assigning it to each collocate
5. Import into iLex	(empty)
6. Evaluation of results	(empty)
7. Large extraction	Repeat the entire process on a large number of lemmas

[220] I would like to thank Dr. Iztok Kosem for helping with the organization of the procedure and the idea for this table.

This dataset in XML format is automatically imported into the database in the iLEX DWS. Figure 9-5 below shows the initial panel when a project is opened in iLEX. As can be seen, on the left-hand side of the page is an index with all extracted lemmas in alphabetical order. Figure 9-5 shows the entry view, while in Figure 9-6 XML view is displayed.

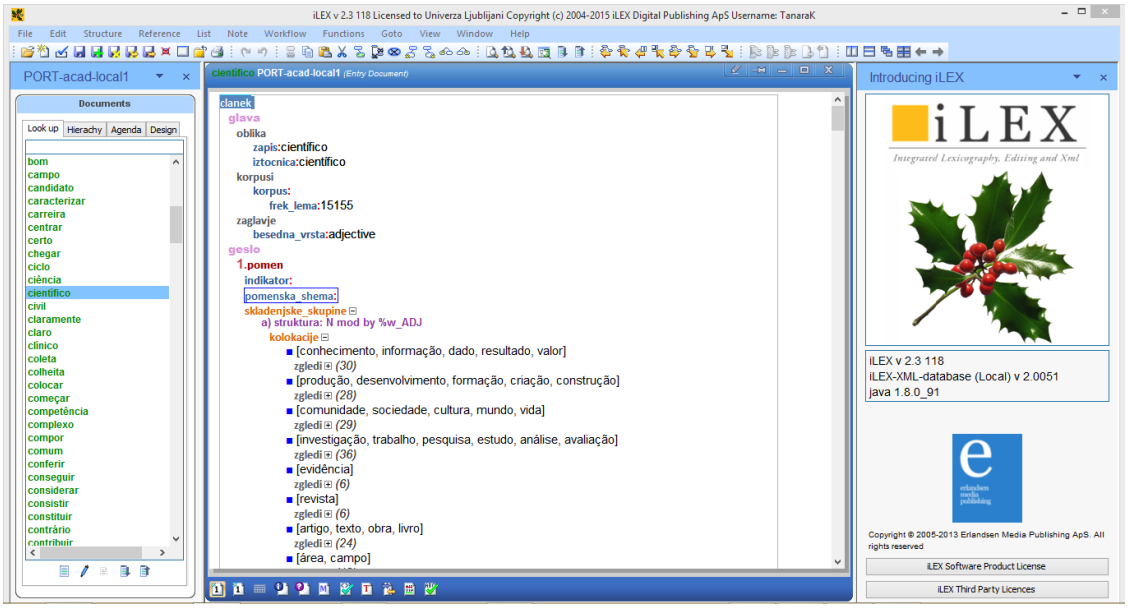


Figure 9-5 Entry view in iLex

Evaluation of the results will be reported in the next section (9.1.3). As to the second extraction, a decision was made to conduct it on a set of sample lemmas rather than a large extraction, as will be shown in section 9.2 below.

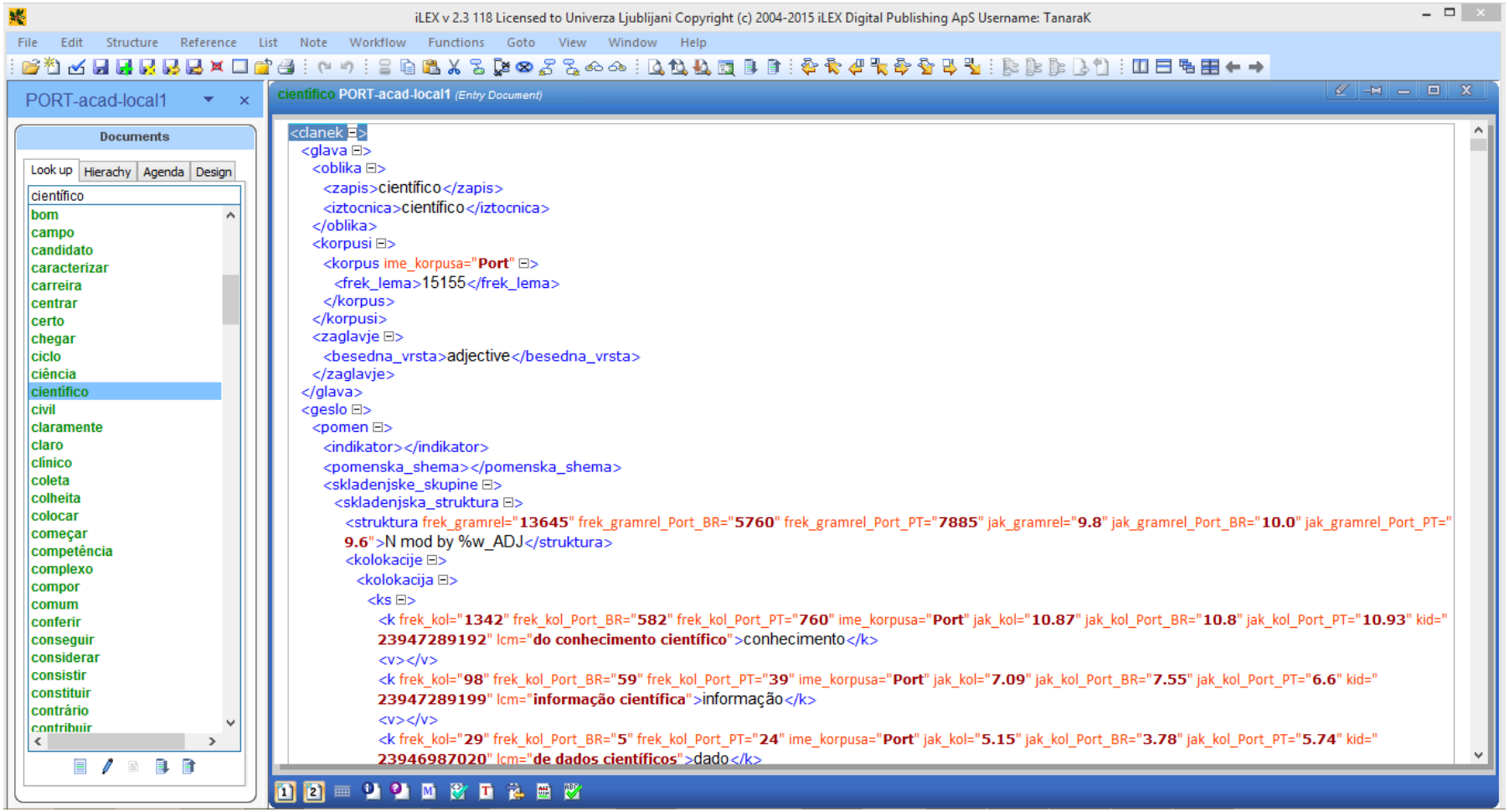


Figure 9-6 XML view in iLex

9.1.3 Evaluation

The evaluation consisted in analysing the automatically extracted data while compiling entries. I paid particular attention to the following points:

a) Missing and problematic grammatical relations
b) Quality of the examples
c) Appropriateness and helpfulness of language variety label assignment
d) General appraisal of the impact of the decisions made for the extraction procedure on the process of entry compilation

As previously mentioned, the process of entry writing with the semi-automated approach begins directly in the DWS. These are the steps that were taken:

1. Remove false collocates, i.e. "those that are incorrectly identified in the word sketch due to reasons such as tagging problems or over-inclusiveness of the grammatical relation" (Gantar et al. 2016, p.19). Guiding question: "Does this collocation reflect the gramrel?
2. Remove false examples, i.e. "those that do not attest the collocation correctly" (Gantar et al. 2016, p.19)
3. Identify senses and subsenses
4. Sort collocates under meanings
5. Devise sense indicators (for sense menu)
6. Semantic frames for adjectives and nouns
7. Identification of compounds
8. Identification of phraseological units
9. Definition writing
10. Definition writing for compounds

As an illustrative example of the initial step, i.e. elimination of false collocates, the figure below (Figure 9-7) shows the disposition of automatically placed collocates under one structure – *%w de+o N* ('career of N') - in the iLex DWS. The keyword (%w) is the noun *carreira* ('career' in the sense presented here).

Figure 9-7 Collocates of *carreira* in iLex. Gramrel: %w de+o N

The collocates for *carreira* in this particular gramrel that were automatically extracted from CoPEP and imported into iLex are, among others, *mulher* ('woman'), *silva* ('silva', a surname), *primeiro-ministros* ('prime-ministers'), and *filho* ('son'). The word *silva* immediately stands out as a non-collocate, as it is a surname. A verification of the examples confirmed that (Figure 9-8).

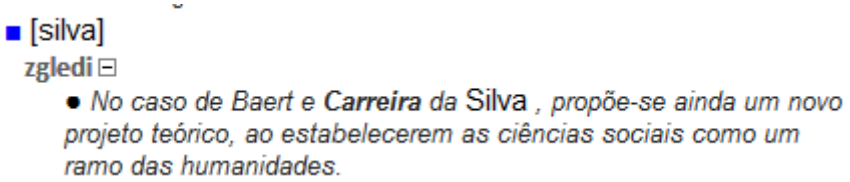


Figure 9-8 False collocate of *carreira*.

In this case, *Carreira* (with capital letter), together with *da Silva*, makes up a typical Portuguese surname, indicating that *carreira da silva* is not a collocation.

9.1.3.1 Sketch grammar

The purpose of the evaluation of grammatical relations was to test the sketch grammar. Missing relations in the automatically extracted data could indicate problems with the queries, while matching more bad collocations than good ones suggested that grammatical relations did not produce relevant results. In both cases, measures should be taken: on the one hand, problematic queries would have to be modified, and on the other, useless relations should be excluded from the settings of subsequent automatic extractions.

One major problem that comparison between the sketch grammar (Portuguese v7.1) and the extracted relations in the entry database revealed was that relations comprising verbs with *-se* were never matched. A detailed examination of the queries,

the tagset, the tagger, and the corpus in order to identify the source of the problem indicated that the option followed by Freeling for tagging these types of verbs, namely, lemma+*se* (e.g., *tornar+se*), was not appropriate, at least for CQL queries matching. As this issue was described in Chapter 7, I will not repeat it here. It is worth highlighting, however, that the solution to this problem involved the development of a new pipeline for Portuguese by the Sketch Engine team. Consequently, I rewrote these relations in the new version of my sketch grammar (AcadPortSkG_v3).

Another example of a revealing result of this evaluation concerns the grammatical relation noun subject of verb. This relation was composed of three queries: one for picking up a singular noun as subject, another for finding plural nouns as subject, and a third one for matching compound subjects. This is how it looked:

```
*DUAL
=%w_N subj of V/N subj of %w_V
1:"NC.P.*"[word!="\."]{0,7}2:"VM.*3P."
1:"NC.S.*"[word!="\."]{0,7}2:"VM.*3S."
1:"NC.[P|S].*""CC""NC.[P|S].*"[word!="\."]{0,7}2:"VM.*3P."
```

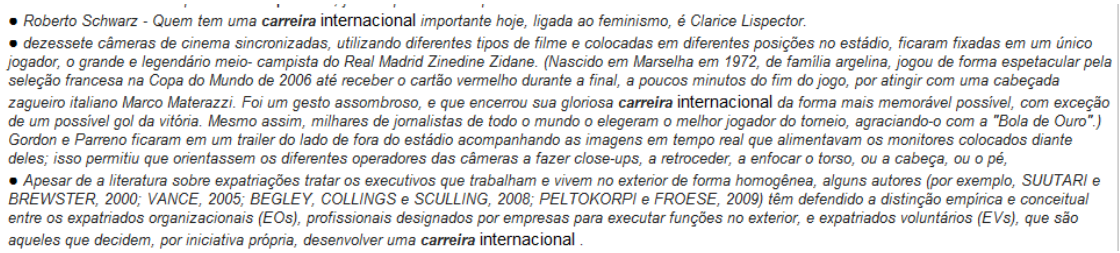
This relation has proven to be particularly problematic as it is highly unstable, given the fact that, in many cases, the noun anticipating the verb is not necessarily its subject. For a language as flexible as Portuguese, where the canonical order is often subverted, the use of a POS-tagged sketch grammar presents a serious challenge.

Nevertheless, as the relationship between subject and verb is an indispensable piece of information in the automatic extraction of data, and considering that the above gramrel does pick up correct information, this relation was kept in the new sketch grammar. For the next extraction, it was decided that the third query (`1:"NC.[P|S].*""CC""NC.[P|S].*"[word!="\."]{0,7}2:"VM.*3P."`) should be eliminated, as it did not yield relevant results.

9.1.3.2 GDEX

It is well-known that it is far from a simple task to find good dictionary examples in a corpus (Atkins and Rundell 2008, p. 457), and this issue becomes even more critical when compiling a lexical database in which many more examples are required. This is why automatic extraction of examples is one of the key features in the semi-automated approach to dictionary-making.

Analysis of the automatically extracted candidates for good dictionary examples revealed some important points that need improvement for the new extraction. One of them concerned sentences boundaries markup, which had been automatically performed on the Sketch Engine. In many cases, text passages containing more than one sentence were identified as one sentence only, as shown in Figure 9-9. Given that each bullet refers to one sentence, it can be seen that the second good dictionary example candidate comprises many sentences.


• *Roberto Schwarz - Quem tem uma* ***carreira*** internacional *importante hoje, ligada ao feminismo, é Clarice Lispector.*

• *dezessete câmeras de cinema sincronizadas, utilizando diferentes tipos de filme e colocadas em diferentes posições no estádio, ficaram fixadas em um único jogador, o grande e legendário meio- campista do Real Madrid Zinedine Zidane. (Nascido em Marselha em 1972, de família argelina, jogou de forma espetacular pela seleção francesa na Copa do Mundo de 2006 até receber o cartão vermelho durante a final, a poucos minutos do fim do jogo, por atingir com uma cabeçada zagueiro italiano Marco Materazzi. Foi um gesto assombroso, e que encerrou sua gloriosa* ***carreira*** internacional *da forma mais memorável possível, com exceção de um possível gol da vitória. Mesmo assim, milhares de jornalistas de todo o mundo o elegeram o melhor jogador do torneio, agraciando-o com a "Bola de Ouro".) Gordon e Parreno ficaram em um trailer do lado de fora do estádio acompanhando as imagens em tempo real que alimentavam os monitores colocados diante deles; isso permitiu que orientassem os diferentes operadores das câmeras a fazer close-ups, a retroceder, a enfocar o torso, ou a cabeça, ou o pé,*

• *Apesar de a literatura sobre expatriações tratar os executivos que trabalham e vivem no exterior de forma homogênea, alguns autores (por exemplo, SUUTARI e BREWSTER, 2000; VANCE, 2005; BEGLEY, COLLINGS e SCULLING, 2008; PELTOKORPI e FROESE, 2009) têm defendido a distinção empírica e conceitual entre os expatriados organizacionais (EOs), profissionais designados por empresas para executar funções no exterior, e expatriados voluntários (EVs), que são aqueles que decidem, por iniciativa própria, desenvolver uma* ***carreira*** internacional .

Figure 9-9 Multi-sentenced good dictionary example candidate

Another aspect that had room for improvement was the GDEX configuration. As described in Chapter 8, the version used for the extraction experiment (Portuguese v7) was a fairly simple adaptation of the GDEX configuration for Slovene (version 2). It seemed plausible to expect that CoPEP-based configuration developments should have a positive impact on the quality of the examples.

A decision was thus made to further develop GDEX configurations for academic Portuguese using statistics from CoPEP. The second extraction makes use of this new version, whose process of development has been fully described in Chapter 8.

9.1.3.3 Language variety

Labels for language variety have proven to be extremely useful for the entry compilation routine. As shown in Figure 9-10, the label facilitates allocation of specific uses of collocations according to the country of use.

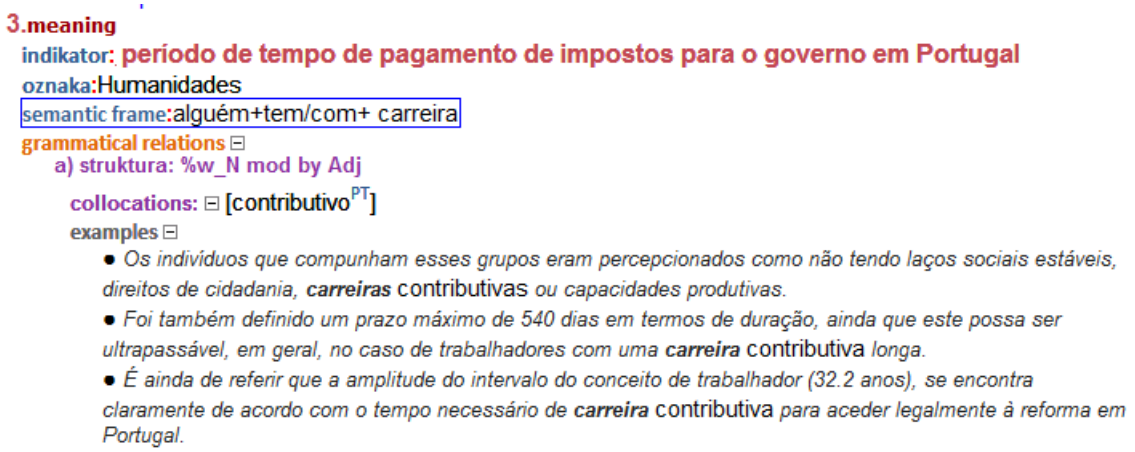


Figure 9-10 *Carreira contributiva*. Collocation used in European Portuguese.

Labels should be even more convenient for lexicographers dealing with normalised corpora of Portuguese. As explained in Chapter 6, with the advent of a new orthographic agreement between the Member States of CPLP (the AO90), spelling rules have been unified, without, however, discarding the traditional lexicography of each country. This means that CoPEP, for instance, is composed of texts following four different coexisting spelling norms. In order to reduce variance, I opted for normalizing my corpus, updating the spelling of words to the current, official rules (see Chapter 6). Thanks to the label assignment function added to the procedure, words with the same spelling (homographs) are analysed according to their occurrences in each language variety subcorpus.

Information on domain (‘*oznaka’*, in Figure 9-10 above) – either great area of knowledge or school – required verification of text types frequency in the corpus (in the Sketch Engine). At the time of the experiment, the Pre-final corpus was not fully annotated with metadata, which were visible only via the file name. In Figure 9-11 below, these are seen in blue at the left end of each concordance line of the collocation *carreira contributiva*. With mouse hover, the full name is displayed and the remaining metadata can be visualised (ISSN number, year, etc.; see Chapter 6).

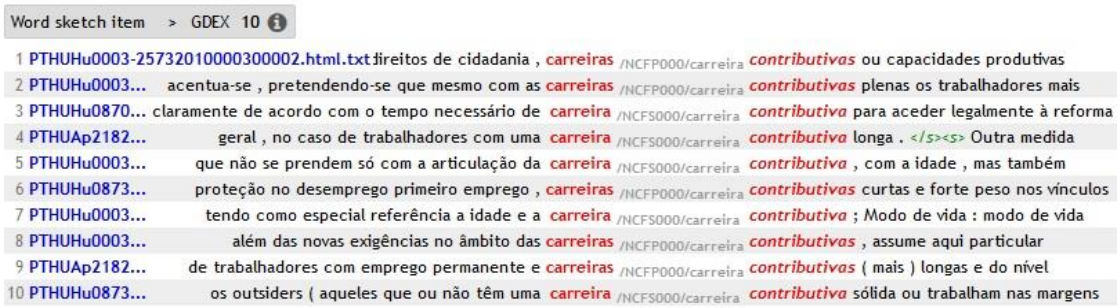


Figure 9-11 *Carreira contributiva* concordance lines with metadata in the file names.

As can be seen, this collocation occurred in the School of Humanities (HU in the file name), with eight occurrences in the Human Sciences great area of knowledge (Hu) and two in the Applied Sciences. Although it is apparent that CoPEP is a fairly small corpus and generalizations cannot be made, for the purposes of DOPU design, I opted to fill out the domain component in the entry so that what is observed in my project is displayed.

9.1.3.4 General appraisal of the procedure

Evaluation of the outcomes indicated that the automatically extracted lexical information could be used as a basis for the writing of entries of the dictionary of academic Portuguese after some editing, namely, elimination of false collocates and examples, together with the addition of information obtained from direct examination of the corpus.

Among the possible areas of improvement in the procedure, it became clear that improved quality of the corpus would bring about a lot of benefits for both the whole process of extraction and DOPU lexical content. I decided to focus on this, rather than try to come up with additions for procedure upgrading, although this is my plan for future works.

9.2 Second extraction

Initially, the plan was to conduct an extraction on the whole candidate headword list (single items only) after carrying out the experiment on a set of sample lemmas. However, as shown above, a more thorough development of resources and tools was deemed necessary, so a full extraction will be left for future work. At this moment, a

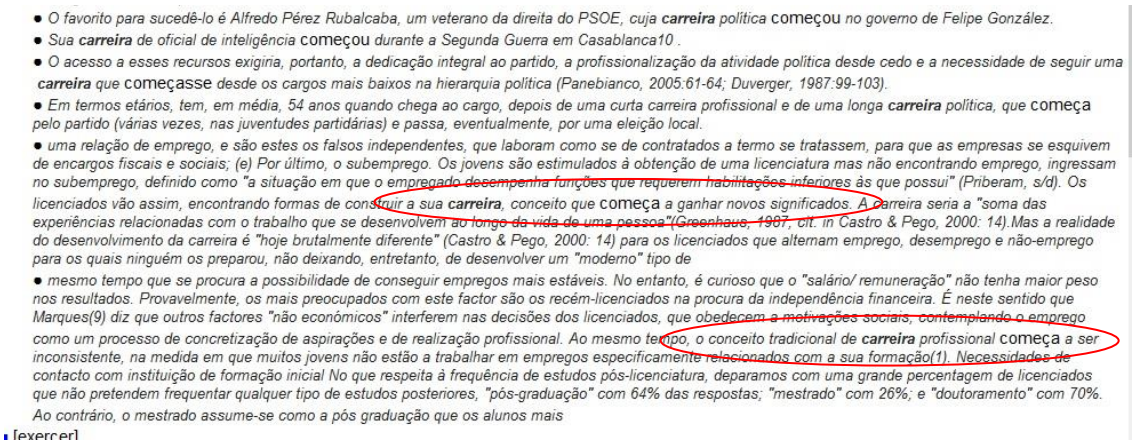
very similar list to the one used in the experiment was applied, facilitating a comparison with this latest version.

In consequence, the differences between the experiment of extraction and the second extraction mostly refer to the refinement of CoPEP, sketch grammar and GDEX configuration, which were described extensively in Part II of this thesis and will not be covered again here.

Analysis of the outcomes of the second extraction confirmed that improvement of the resources and tools enhanced the quality of the lexical content extracted. For example, let us compare information extracted from CoPEP and imported into iLex between the experiment and the second extraction.

a) Relation noun subject of verb. Noun: *carreira*, 'career':

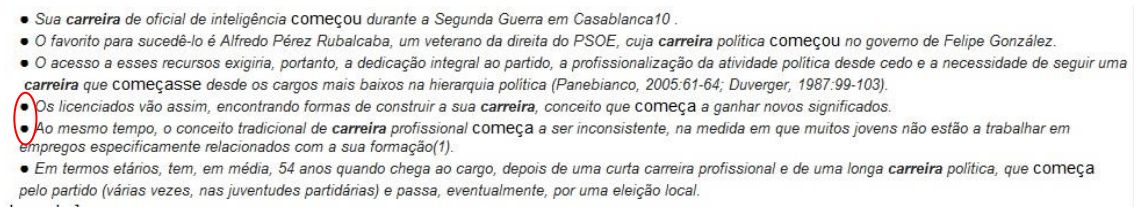
For the collocation *carreira* + *começar* ('to start', verb), Figure 9-12 shows the information extracted in the experiment, while Figure 9-13 comprises examples from the second extraction.



- *O favorito para sucedê-lo é Alfredo Pérez Rubalcaba, um veterano da direita do PSOE, cuja* ***carreira*** *política* começou *no governo de Felipe González.*
- *Sua* ***carreira*** *de oficial de inteligência* começou *durante a Segunda Guerra em Casablanca10 .*
- *O acesso a esses recursos exigiria, portanto, a dedicação integral ao partido, a profissionalização da atividade política desde cedo e a necessidade de seguir uma* ***carreira*** *que* começasse *desde os cargos mais baixos na hierarquia política (Panebianco, 2005:61-64; Duverger, 1987:99-103).*
- *Em termos etários, tem, em média, 54 anos quando chega ao cargo, depois de uma curta carreira profissional e de uma longa* ***carreira*** *política, que* começa *pelo partido (várias vezes, nas juventudes partidárias) e passa, eventualmente, por uma eleição local.*
- *uma relação de emprego, e são estes os falsos independentes, que laboram como se de contratados a termo se tratassem, para que as empresas se esquivem de encargos fiscais e sociais; (e) Por último, o subemprego. Os jovens são estimulados à obtenção de uma licenciatura mas não encontrando emprego, ingressam no subemprego, definido como "a situação em que o empregado desempenha funções que requerem habilitações inferiores às que possui" (Priberam, s/d). Os licenciados vão assim, encontrando formas de construir a sua* ***carreira****, conceito que* começa *a ganhar novos significados. A carreira seria a "soma das experiências relacionadas com o trabalho que se desenvolvem ao longo da vida de uma pessoa"(Greenhaus, 1987, cit. in Castro & Pego, 2000: 14).Mas a realidade do desenvolvimento da carreira é "hoje brutalmente diferente" (Castro & Pego, 2000: 14) para os licenciados que alternam emprego, desemprego e não-emprego para os quais ninguém os preparou, não deixando, entretanto, de desenvolver um "moderno" tipo de*
- *mesmo tempo que se procura a possibilidade de conseguir empregos mais estáveis. No entanto, é curioso que o "salário/ remuneração" não tenha maior peso nos resultados. Provavelmente, os mais preocupados com este factor são os recém-licenciados na procura da independência financeira. É neste sentido que Marques(9) diz que outros factores "não económicos" interferem nas decisões dos licenciados, que obedecem a motivações sociais, contemplando o emprego como um processo de concretização de aspirações e de realização profissional. Ao mesmo tempo, o conceito tradicional de* ***carreira*** *profissional* começa *a ser inconsistente, na medida em que muitos jovens não estão a trabalhar em empregos especificamente relacionados com a sua formação(1). Necessidades de contacto com instituição de formação inicial No que respeita à frequência de estudos pós-licenciatura, deparamos com uma grande percentagem de licenciados que não pretendem frequentar qualquer tipo de estudos posteriores, "pós-graduação" com 64% das respostas; "mestrado" com 26%; e "doutoramento" com 70%. Ao contrário, o mestrado assume-se como a pós graduação que os alunos mais*

[exercer]

Figure 9-12 Examples of the collocation *carreira* + *começar*. Extraction experiment.



- *Sua* ***carreira*** *de oficial de inteligência* começou *durante a Segunda Guerra em Casablanca10 .*
- *O favorito para sucedê-lo é Alfredo Pérez Rubalcaba, um veterano da direita do PSOE, cuja* ***carreira*** *política* começou *no governo de Felipe González.*
- *O acesso a esses recursos exigiria, portanto, a dedicação integral ao partido, a profissionalização da atividade política desde cedo e a necessidade de seguir uma* ***carreira*** *que* começasse *desde os cargos mais baixos na hierarquia política (Panebianco, 2005:61-64; Duverger, 1987:99-103).*
- *Os licenciados vão assim, encontrando formas de construir a sua* ***carreira****, conceito que* começa *a ganhar novos significados.*
- *Ao mesmo tempo, o conceito tradicional de* ***carreira*** *profissional* começa *a ser inconsistente, na medida em que muitos jovens não estão a trabalhar em empregos especificamente relacionados com a sua formação(1).*
- *Em termos etários, tem, em média, 54 anos quando chega ao cargo, depois de uma curta carreira profissional e de uma longa* ***carreira*** *política, que* começa *pelo partido (várias vezes, nas juventudes partidárias) e passa, eventualmente, por uma eleição local.*

Figure 9-13 Examples of the collocation *carreira* + *começar*. Second extraction.

Immediately visible is the difference in example lengths, confirming that the sentences boundaries markup has been fixed. Indeed, examples 4 and 5 in the second extraction are contained in the very large text excerpts that had been extracted as good example candidates in the experiment. By stripping off what was unnecessary (extra sentences), the system kept only the matching patterns. The fact that the examples are the same indicate that modification of the GDEX configuration did not have a high impact,[221] although the order of the examples did change.

b) New grammatical relation: symmetric *e|ou*

This gramrel was not present in the experiment. However, as can be seen from the examples in Figure 9-14, these are interesting collocations that contribute to sense identification. The keyword was the verb *começar* (‘to start’), with two collocates that tend to be considered synonyms, *terminar* (‘to finish’) and *acabar* (‘to end’).

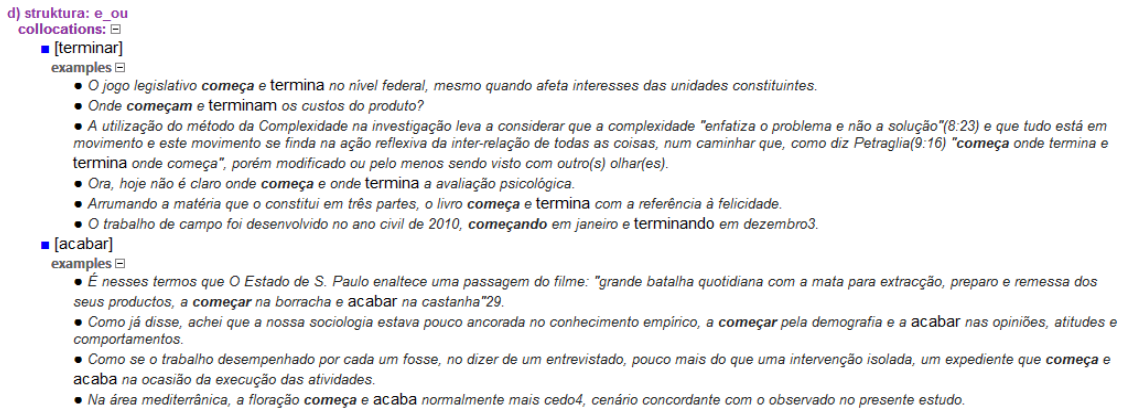


Figure 9-14 Symmetric relation *e|ou* (keyword= *começar*; collocates= *terminar* and *acabar*).

The verb *iniciar* (‘to begin’) was not in the set of extracted sample lemmas, but it would definitely be useful to compare these collocations, namely, *começar* and *terminar* versus *iniciar* and *terminar*, for instance, to be better able to identify their senses. Of course, it is possible to resort the corpus in the Sketch Engine for analysis of these structures’ behaviour.[222] However, as the evaluation here refers to the extraction results, this is to be used later when compiling entries.

[221] In fact, the new GDEX configuration for academic Portuguese was not strikingly different from the first version (Portuguese v7), which was used in the experiment. See Chapter 8.

[222] Once the analysis goes back to the Sketch Engine, the Sketch Diff function can provide insightful additional information.

With improved resources and tools and better extracted lexical content, future steps concern upgrading the procedure. One of the questions that will deserve dedicated research concerns the evaluation of the influence of areas of knowledge on the efficiency of GDEX configuration classifiers: they might require specific values or even new classifiers.

For this design, I am already satisfied with the quality of the extracted data. After all, the procedure has shown that such a method can indeed be applied for DOPU with very good results.

9.3 Summary

Continuation of the development of the procedure as well as resources and tools aiming to enhance the quality of both the process of automatic extraction of data, as well as the final product, i.e. the dictionary, would most certainly be ideal. However, this is an ambitious enterprise. Thus, it is not surprising that a dedicated team of lexicographers, linguists and computational linguists, as well as proper allocation of time and, very importantly, funding, are of utmost need. As these conditions cannot be met now, an experiment at an intermediary stage of the research and a second extraction were performed in order to demonstrate the potentiality of adoption of this approach for Portuguese lexicography.

Chapter 10 Macrostructure

As previously stated in Chapter 4, in this thesis, macrostructure consists of the A-Z entry list, the front matter and the back matter (cf. Landau, 2001). Front and back matters are terms used in printed dictionaries to refer to additional references presented before and after the headword list, respectively. In an online dictionary, front and back matters evidently no longer denote the location of information in the book, but still indicate the inclusion of other materials in a dictionary besides the word list proper. The focus in this chapter, however, will be on the headword list (10.1), with only a brief description of potential contents of supplement material (10.2).

Hence, the main purpose of this chapter is to present the process of candidate headword list building, focussing on a number of decisions of theoretical and methodological nature that had to be made and were pointed out in Chapter 4. This means that, instead of providing a closed set of headwords, this chapter intends to bring forth principled choices deriving from CoPEP_AO90's data handling.

10.1 The headword list

In Chapter 4, two basic conditions for the compilation of a DOPU candidate headword list were presented and properly justified: that the headword list should be built from CoPEP and that frequency should be the main criterion for candidate headword selection. Other factors were pointed out as fundamental for decision-making and will be tackled in the following subsections.

For organizational purposes, I will begin by touching upon issues related to the delimitation of vocabulary in the headword list (10.1.1), to then move on to specific matters at lexical entry level (10.1.2).

10.1.1 Delimitation of vocabulary

In Chapter 4, it was defined that frequency should be the main criterion for the determination of what words should be included in the candidate headword list of DOPU. Another factor had to be assessed that affects the delimitation of the vocabulary

in the candidate headword list: VOC ('Common Orthographic Vocabulary of the Portuguese Language'). I will start with corpus frequency to then touch upon VOC.

10.1.1.1 Corpus frequency

Selecting headwords based on corpus frequency is a common practice nowadays (de Schryver, 2013). Nonetheless, most authors of dictionaries and theoretical lexicographers do not mention cut-off point values (cf. Welker, 2004, pp. 93-95). Among those few who do approach the issue, I refer here to Biderman (1984b,1996b) and Kosem (2010) as suitable sources for the determination of minimum frequency in DOPU.

Biderman (1984b) argues that a dictionary of Brazilian Portuguese for elementary school students would probably need no more than 5,000 entries, thus proposing a minimum frequency of 5 in a 5-million-word corpus. In Biderman (1996b), the author resorted to lexical-statistic measures employed in two corpus-based projects as a reference for her project on the *Fundamental Vocabulary of Brazilian Portuguese*. The first study had been developed for the European Portuguese version of the *Fundamental Portuguese – Vocabulary* , which was carried out at the University of Lisbon in the 1970s. A formula created by Paul Rivenc established a minimum frequency of 40, considering word frequency and dispersion. The other measure had been calculated by John B. Carroll for the American Heritage Word Frequency Book, which suggested a cut-off point of 20 for a 5 million-word corpus of written English texts. Biderman opted for the later value for her research.

More recently, Kosem (2010, p.171) argued when developing his model for a dictionary of academic English that vocabulary coverage should be comprehensive in order to meet university students' reading and writing needs. He thus compiled a corpus-driven lemma list to serve as the basis for selection of a candidate headwords list. However, the full list was not used; rather, a minimum frequency limit of 5 (i.e. $\leq$ 4) was set in order to avoid very rare lemmas.

Despite the evident differences between the projects of these lexicographers, both of them draw attention to the fact that *hapax legomena* (words occurring only once) should be discarded as they reflect idiosyncratic language use, which is just the opposite of what a dictionary aims to describe.

I decided to experiment the values mentioned above, namely, 40, 20, 5, and 0, as minimum frequency limits in CoPEP_AO90 in order to define a suitable threshold for the DOPU lemma list. Table 10.1 below shows the results of the total number of lempos[223] and number of lempos per word category in CoPEP_AO90, while in Table 10.2, data on *hapax legomena* can be visualised. It should be highlighted that a lempos list, rather than a lemma list, was used for these calculations as it enables breaking down the list into word classes. Moreover, the automatic extraction of data demonstrated in Chapter 9 was based on the lempos list.

Table 10.1 Minimum cut-off values and the number of lempos in CoPEP_AO90, broken down into word classes

Cut-off frequency value	Total	Number of verbs	Number of nouns	Number of adjectives	Number of adverbs
40	25,152	2,849	16,991	4,313	662
20	38,453	3,848	27,068	6,186	881
5	92,611	7,085	69,481	12,887	1,681
0	437,466	30,619	300,128	45,766	7,701

Table 10.2 Total number of *hapax legomena* in CoPEP_AO90, broken down into word classes

	Hapax legomena
Total	252,765
Number of verbs	18,005
Number of nouns	161,037
Number of adjectives	22,619
Number of adverbs	4,684

As can be seen, *hapax legomena* make up 57.71% of the corpus. A quick overview of this end-list tail reveals tokenization problems and irregular characters (e.g., email addresses, parts of formulas, foreign words, misspelt words, etc.), which are

[223] Lempos stands for lemma with POS-tag indication, as in *estudo*-n (‘study’; -n, noun). See Chapter 5 for further information.

unwanted in the DOPU database (see Appendix F in the CD-Rom for the full lempos list).

Although CoPEP is smaller than the corpus that Kosem created for compiling the *Dictionary of Academic English*, as my project is also aimed at university students, I decided to adopt the same threshold of 5, thus using a 92,611-lempos list as a departure point for candidate headword building.

Nevertheless, attention should be drawn to the fact that the resulting list could not be simply considered the candidate headword list. This is due to at least three reasons. Firstly, there have been tokenization errors and unwanted strings of characters occurring above the threshold limit. Secondly, this lempos list comprises single items, while the DOPU headword list will also include multi-word items (as will be discussed below). Thirdly, lemmatization errors (some word-forms have not been lemmatized, so, for instance, a plural and a singular form were identified as two lempos, instead of being transformed into a lemma) indicated the need of item conflation under one lemma.

10.1.1.2 The official vocabulary of Portuguese (VOC)

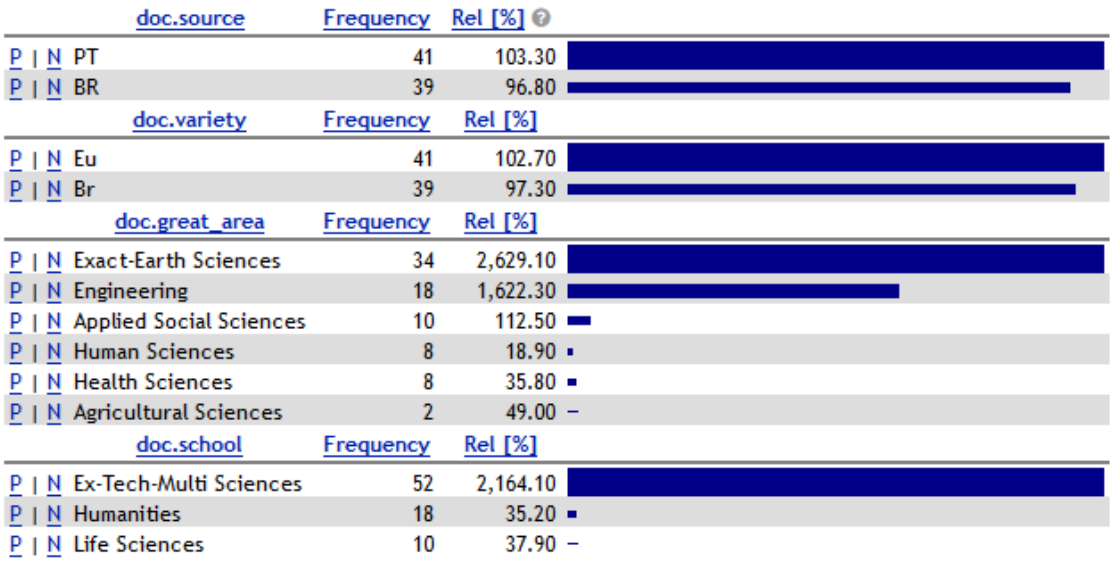
Throughout the 20th century, efforts had been made to regulate the orthography of the Portuguese language both in Brazil and Portugal. However, the rules were divergent in many aspects, resulting in the coexistence of different norms governing one language. The *Acordo Ortográfico da Língua Portuguesa* of 1990 (AO90) ('The Portuguese Language Orthographic Agreement') (see Chapter 1 and 6) was created to unify and simplify the orthographic rules between BP and EP, without overriding lexicographical tradition of each variety. In 2015, six years after the actual implementation of the treaty, the Common Orthographic Vocabulary of the Portuguese Language (VOC) was launched as the CPLP-ratified lexicographical tool gathering the different official vocabularies of each national variety in one place.

As a document of authority, VOC governs the orthography of each national variety, stating how words should be spelt and presenting accepted variants. In other words, it defines the orthographic norm that must be followed.

The question, in the case of DOPU, is whether VOC should be taken as a measure for the definition of inclusion and exclusion of words. More specifically,

should words that are not attested in VOC be eligible for headword status? An illustrative case is shown below.

The noun *sequenciamento* (approximately 'sequencing') occurs in CoPEP_AO90 with a total frequency of 94, 80 times in the singular form and 14 times in the plural form. A corpus analysis in the Sketch Engine reveals that it has almost the same frequency in both BP and EP subcorpora (Figure 10-1), when used in the singular form, while all the 14 occurrences of the plural form are found in one paper written in EP (Exact-Earth Sciences).



doc.source	Frequency	Rel [%]
P | N PT	41	103.30
P | N BR	39	96.80
doc.variety	**Frequency**	**Rel [%]**
P | N Eu	41	102.70
P | N Br	39	97.30
doc.great_area	**Frequency**	**Rel [%]**
P | N Exact-Earth Sciences	34	2,629.10
P | N Engineering	18	1,622.30
P | N Applied Social Sciences	10	112.50
P | N Human Sciences	8	18.90
P | N Health Sciences	8	35.80
P | N Agricultural Sciences	2	49.00
doc.school	**Frequency**	**Rel [%]**
P | N Ex-Tech-Multi Sciences	52	2,164.10
P | N Humanities	18	35.20
P | N Life Sciences	10	37.90

Figure 10-1 Text types frequency of *sequenciamento* in CoPEP_AO90

Sequenciamento is included in the three reference dictionaries of BP used in this thesis, namely, Aurélio, Borba, and Houaiss, but is not in the nomenclature of any of the dictionaries of EP, that is, Academia, Infopedia, Porto and Priberam. Moreover, *sequenciamento* also occurs in the *Corpus Brasileiro* (1.006), Portuguese Web 2011 (BP:1406; EP: 106; unknown: 06); CRPC (PE:01); *Projeto AC/DC*- Linguateca (909). Nevertheless, *sequenciamento* is not attested in VOC.[224]

As can be seen, this issue is far from simple. It requires that matters related to the norm and the role of DOPU be brought into play. Hence, this topic should be discussed by the editorial team when the DOPU project is implemented. For the design,

[224] That is, not in the Orthographic Vocabulary of Portuguese -VOP (Portugal) nor in the Orthographic Vocabulary of the Portuguese Language -VOLP (Brazil).

it was decided that cases like this should be considered candidates if they meet the other criteria presented in this chapter.

10.1.2 Lexical entries

The factors approached above are complemented with decision-making at the level of lexical entries. Drawing on the contributions of Atkins (1992/3, 2008), Atkins and Rundell (2008), Gouws (2003), Landau (2001), and Svénsen (2009), to name but a few lexicographers of reference, the following aspects must be taken into account for the definition of candidate headword lists.

Table 10.3 Decisions concerning headword status

Types of words	1. ***Which common words*:** a. **Abbreviations and contractions?** b. **Partial words (prefixes, suffixes)?** c. **Multiword expressions (idioms, compounds, fixed and semi-fixed phrases, collocations)?** d. **Inflected forms (irregular comparative and superlative of adjectives, verb inflections)?** e. **Derived forms?** 2. ***Which proper names*:** f. **Place names (oceans, countries, states, capital cities, mountains, regions, etc.)?** g. **Personal names (people's names: theorists, related adjectives - *Bakhtinian*? ethnic groups, etc.)** h. **Other names (ceremonies, institutions, languages, trademark, religions, etc.)?**
Homonymy X polysemy	Words with the same spelling but different meanings: homonymy (different headwords) or polysemy (one headword)?
Variant forms	Variant forms within one language variety?
Loan words	Latin words/expressions? Words from other languages?
Miscellaneous	Numerals? Formulae? Symbols?

10.1.2.1 Type of words

10.1.2.1.1 Common words

a) Abbreviations and contractions

It has been decided that abbreviations and contractions are included in the candidate list as they are frequent and pervasive in the whole corpus. Figure 10-2 shows

some of the automatically identified acronyms in the corpus, while Figure 10-3 gives an overview of the distribution of abbreviations across the six areas of knowledge.

P | N	ONU	1,799
P | N	M	1,795
P | N	SUS	1,732
P | N	PT	1,700
P | N	US	1,689
P | N	S	1,612
P | N	HIV	1,560
P | N	T	1,534
P | N	IMC	1,462

Figure 10-2 Abbreviations in CoPEP

	doc.great_area	Frequency	Rel [%]
P | N	Human Sciences	322,625	80.70
P | N	Health Sciences	257,691	122.00
P | N	Applied Social Sciences	80,227	95.40
P | N	Agricultural Sciences	64,081	166.10
P | N	Exact-Earth Sciences	17,261	141.10
P | N	Engineering	14,789	140.90

Figure 10-3 Distribution of abbreviations across areas of knowledge of CoPEP

b) Partial words

Following Kosem (2010), prefixes and suffixes will be part of DOPU, as well as non-autonomous units, such as Greek and Latin roots. Although it is known that a word formed from a process of derivation and composition will have its particular characteristics of use and acquire specific senses, the provision of semantic information is believed to be beneficial.

c) Multi-word expressions

The importance of tackling multi-word expressions at the macrostructure level of DOPU is due to their status as headwords. While hyphenated compounds and MWEs covering one lexical unit are considered candidates for inclusion in the nomenclature, identification of some other types of MWEs is difficult and requires careful lexical analysis. This is why a definition of those MWEs that are eligible for headword status cannot be done at this stage of the research.

Some theoretical works that might serve as guidelines to help lexicographers to make decisions when involved in the analysis of MWEs for DOPU are Biderman (2005) for detailed account of different constructions, identification tests, and a number of

examples; and Correia and Lemos (2009), who present a concise and example-rich review of compounds. It should be highlighted that both works refer to the Portuguese language.

d) Inflected forms

It is apparent that some inflected forms which clearly make up a new word category – as in the case of participle forms of the verbs, which are very often used as adjectives – will be considered candidates for the headword list. However, such determination cannot be given before actual corpus analysis, when lexicographers will have a chance to examine the forms that are most frequently used and to what extent senses differ. On the other hand, number, gender and degree inflections should not be given headword status, unless the inflection of the lemma constitutes a new lexical unit. Attention should be drawn to the fact that there will be a reference system that will link any inflected word looked up in the search box with its canonical form.

e) Derived forms

In the case of adverbs formed from the process of deadjectival derivation, i.e. those adverbs ending in *-mente* ('-ly'), these will be given headword status. This is due to the fact that, although the process of derivation is stable, namely, these adverbs are formed by the addition of the feminine form of the adjective followed by *-mente*, their meanings vary according to the senses (lexical units) of the base, as well as their context. That is, each adverb formed from the same base covers different lexical units.

f) Proper names:

Place names: oceans, countries, states, capital cities, mountains, regions, etc.; people's names: theorists, related adjectives (e.g. Hegelian), ethnic groups; names for ceremonies, institutions, languages, trademarks, religions, etc. will all be given candidate headword status.

10.1.2.1.2 Homonymy X polysemy

Polysemy is when one lexical item has more than one meaning, while homonymy concerns different lexical items sharing a lexical form, but having two very distinctive meanings. As mentioned in Chapter 4, efforts have been made to try to establish precise criteria for the determination of this distinction, but there is no consensus (Landau, 2001, p. 100). For lexicography, this issue has consequences on the

macrostructure of the dictionary, since homonyms should be given headword status, and on the microstructure, in reference to the treatment of senses in polysemic items.

Traditionally, etymology plays an important role in distinguishing the two phenomena, where words with different etyma are considered homonyms (e.g., Jackson, 2002). Nevertheless, this solution has been widely contested (e.g., Atkins & Rundell, 2008; Biderman, 2000; Landau, 2001; Werner, 1982); with the basic argument revolving around the fact that this issue should be approached synchronically, as users have lost the reference to the origin of the words. Instead, sense relatedness has been given priority, without successfully providing a definitive criterion, though. This is because, according to Koskela (2015), resorting to the argument put forward by Tuggy (1999), "what counts as a semantic relationship is notoriously difficult to define objectively. Intuitive judgements of semantic relations are necessarily subjective and may be influenced by one's inclination of finding differences or similarities in meaning" (Koskela, 2015, p. 459).

Other criteria often adopted for distinguishing between homonyms and polysemic items take into account formal characteristics of lexical items, namely, pronunciation, spelling, and morphosyntactic properties (Cowie, 2001; Dobrovoljc, 2017; Koskela, 2015).

One commonly adopted approach for dictionary-making favours the lexical form of the item in question, including in one entry all the different senses of a headword, irrespective of the nature of these differences (see Haensch, Wolf, Ettinger, & Werner, 1982; Landau, 2001). In other words, homonymy and polysemy are not distinguished for macrostructural decision purposes, being dealt with at the microstructural level. Collins Cobuild English Dictionary and Longman Dictionary of Contemporary English are examples of renowned dictionaries that do not distinguish homonymy and polysemy (Moon, 2000; Moerdijk, 2003).

A decision was made to adopt a polysemic solution for the nomenclature in DOPU and a menu with the presentation of alternative word classes and mini-definitions to help users to choose the meaning searched. One source of inspiration for

such a menu is the one employed in the dictionary of the online lexicographical resource for English called Vocabulary.com,[225] as shown in Figure 10-4 and Figure 10-5 below.

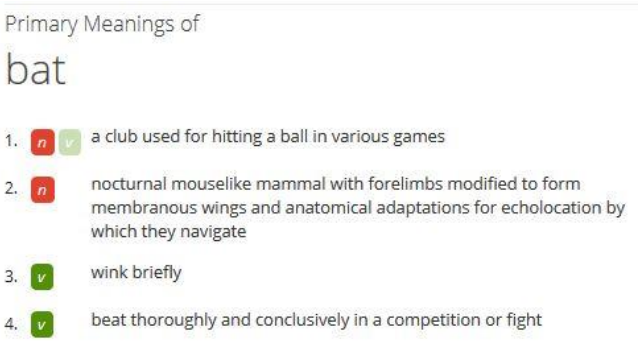


Figure 10-4 Menu for 'bat' (focus on noun) in Vocabulary.com

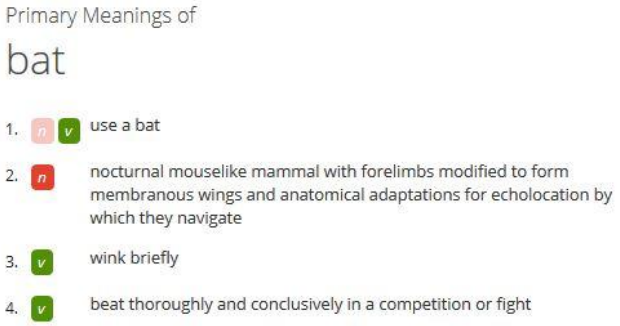


Figure 10-5 Menu for 'bat' (focus on verb) in Vocabulary.com

As can be seen, the lemma 'bat' is represented as being polysemic. The different meanings are presented with their respective word classes, and the colourful POS icons are clickable, thus, when a certain sense and POS is clicked on, automatic scroll down takes the user directly to the desired meaning. Attention should be drawn to the first meaning. In Figure 10-4, the mouse cursor was hovering over the noun icon, making it fully coloured and greying out the colour of the other part of speech, in this case, verb. Simultaneously, as the noun icon was highlighted, a mini-definition was displayed. When moving the cursor over the verb icon (Figure 10-5), its colour was highlighted, while the noun icon got greyed out. Additionally, instead of the mini-definition for the noun meaning of 'bat', the one for the verb can be seen.

[225] https://www.vocabulary.com/dictionary/

I will not debate the adequacy of the mini-definitions presented here nor the meaning coverage given to the lemma as this depends on the purpose of the dictionary and the target-user profile; moreover, it is out of the scope of this section. The point to be made is that there are visually attractive, creative and simplified ways to tackle a polysemic perspective in an online dictionary. There are many more ideas that can be explored and effectively applied if a programmer and a web designer are part of the team.

10.1.2.1.3 Variant forms

The Portuguese language, as explained earlier, is a pluricentric language with eight different varieties. DOPU design covers Brazilian Portuguese and European Portuguese, which have significant differences at lexical, morphological, syntactic, pragmatic and phonetic levels. In consequence, DOPU will be a dictionary of Portuguese that will cater for such differences. Although in the past the enterprise of making one dictionary of Portuguese covering different varieties would be frowned upon, with the advent of electronic lexicography such opposition no longer stands.

The fact that DOPU will be an online dictionary allows different parts of the database to be summoned according to the variety searched by the user, following a similar system to that created for VOC, which uses the MorDeBe relational database (Ferreira, Barbosa, & Janssen, 2008). For building the candidate headword list for DOPU, covering BP and EP entails making a list comprising lemmas that are common to both varieties and lemmas that are exclusive of each variety. In fact, the procedure proposed in Chapter 9 was planned in a way to account for these differences. Labels indicating variety and typical use of variety were assigned at grammatical relation and collocate levels.

Thus, each language variety variant will be given equal headword status, meaning that type/token distinctions, based on the definition of one standard form, will not be adopted.

One additional point to be considered refers to those words whose variants[226] do not occur in CoPEP. In such cases, there should be information in the entry informing

[226] Within one national variety only or equally occurring in both BP and EP. A typical example is the variant forms of the adjective *loiro*/*louro* (‘blond’), which occur in BP and EP.

that there is also variant X, however, it does not occur in CoPEP. This strategy is a sensible solution for situations in which the user searches for the non-occurring variant. Instead of getting an empty result, the user is presented the complete entry of the preferred variant and properly informed about this linguistic fact.

10.1.2.1.4 Loan words and miscellaneous

As shown in Table 10.3 above, this category of analysis for the selection of words with headword status concerns mostly Latin words/expressions, words from other languages, numerals, formulae, and symbols. Given that academic texts from different areas of knowledge undoubtedly make great use of these lexical items, they will all be taken into consideration for building the candidate headword list.

10.1.3 Additional features

Many dictionaries display a partial view of an alphabetic list containing the headword, resembling what the user would see in a print dictionary page (or pages). In the electronic version of *Houaiss* (2009) and *Aurélio Digital* (2010), such lists are shown on the left-hand side of the page (Figure 10-6).

Figure 10-6 Page disposition of alphabetical word list in the electronic versions of Houaiss (2009) and Aurélio (2010)

However, there is another alternative that displays related words instead of neighbouring words due to alphabetical order. MacMillan adopts such an approach, as can be seen in Figure 10-7.

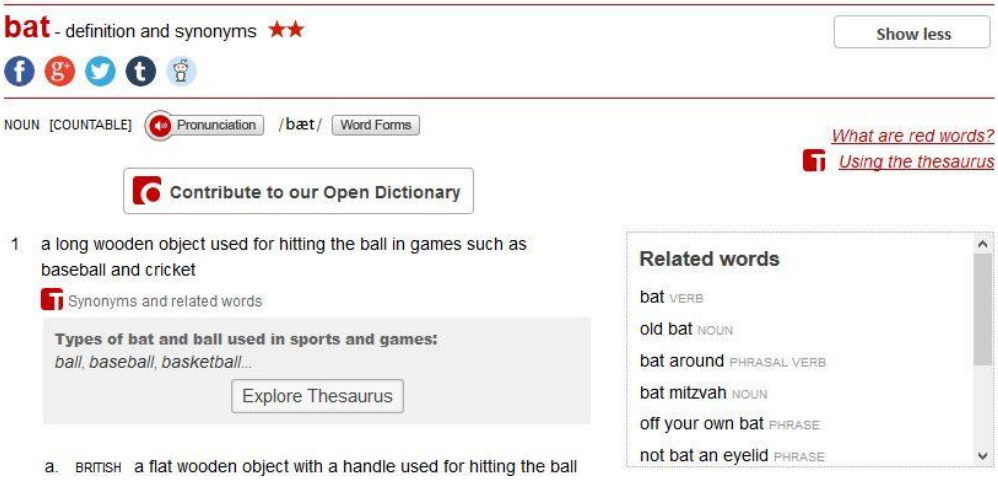


Figure 10-7 Partial view of the entry 'bat' in MacMillan online[227]

Related words, displayed at the right-hand side of the screen, shows words and multi-words expressions that are orthographically related to the entry.

The decision on whether to display an alphabetic list, a related words list, or simply no additional list at all must be taken at the editorial team level. One of the most important conditions for accompanying lists to be included refers to having the technology and the personnel for implementing either lists. Since this design was developed without supporting personnel, this topic will have to be tackled when the project is implemented.

10.2 Supplementary materials

Some of the references that used to be displayed outside the A-Z list in print dictionaries, i.e. in the front and back matters of the dictionary, like abbreviations, proverbs, symbols, etc., can be perfectly included in the headword list in online dictionaries. In a corpus-driven dictionary such as DOPU, the condition for an item to be included is frequency. Other sorts of information, such as a user guide, contributors,

[227] http://www.macmillandictionary.com/dictionary/british/bat_1

and an introduction to the dictionary, can still be displayed via hyperlinks that open up a new page.

One of the advantages of the electronic format of DOPU is that it can contain some other relevant references for university students without interference with the basic nomenclature. For instance, in dictionaries for advanced English learners, it is a common practice to include academic writing tutorials. Although this is outside the lexicographic scope of a dictionary, I understand that it would be useful for students to have access to such material in the same interface as the dictionary. However, as this is beyond the scope of this thesis, the inclusion of this accompanying material shall be discussed in the implementation phase of the project.

10.3 Summary

The principles presented in this chapter are a tentative plan to be further developed within the scope of a long-term project with a well-equipped team of lexicographers, computational linguists, web designers, and corpus linguists. Thus, it is expected that during the implementation of DOPU's lexicographic project, the factors raised here about the candidate list will be reviewed, and suggestions for the addition of new entries and the elimination of others will be brought up for discussion.

Dealing now, at the design stage, with crucial matters that require not only grounding in theory, but also data handling, is expected to ensure a sound and consistent basis for actual DOPU candidate headword building when real implementation of the lexicographic project takes place.

Chapter 11 Microstructure

This chapter aims to present a series of decisions made for the microstructure of the entries, bearing in mind the characteristics of DOPU and its target users.

I have followed the guideline proposed by Atkins (2008 [1992/3], pp. 41-48). According to the author, microstructural decisions comprise:

a) Presentation: how to handle lexical units in the dictionary:
   - Sense differentiation: headword divided into senses
   - Sense order: the order of this senses in the dictionary
b) Data types: what lexically relevant information to record for each lexical unit.

Attention should be drawn to the fact that, despite being displayed here separately, presentation and data types are two aspects of the microstructure that are intricately related; the division here is for organizational purposes only.

11.1 Presentation

Presentation of the lexical units involves how the senses will be differentiated (*sense differentiation*) and in which *order* those senses will be set out in the entry.

11.1.1 Sense differentiation

Atkins and Rundell (2008, p. 296) state that the primordial task of lexicographers when creating entries is "to analyse word forms into meanings, or LUs[228]", which is also usually referred to in lexicography as Word Sense Disambiguation[229] (WSD).

Since the Cobuild project, the canonical manner of identification and distinction of senses in words has relied on analysis of KWIC concordance lines. Nevertheless, the advent of the word sketch function in the Sketch Engine (see Chapter 5 for more) has dramatically changed lexicographic work by taking word sketches as the starting point

[228] Lexical units, i.e. the union of a lexical form and a single sense (Cruse, 1986, p. 77). See Chapter 4.

[229] Although in NLP the term is also used, Atkins and Rundell (op.cit.) explain that, in lexicography, word sense disambiguation means exactly that the context - external (text type, domain) and internal (collocations) - disambiguates the meaning of a word.

of lexical analysis. According to Rundell and Kilgarriff, the word sketch function has proven to work as an accurate, automated sense identifier. The authors (2011, p. 269) explain:

> each of a word's different meanings is associated with particular collocations, so the collocates listed in the word sketches provided valuable prompts in the key task of identifying and accounting for all the word's meanings in the entry.

As presented earlier (see Chapter 9 for a full account), automatic extraction of data from the corpus and import into the dictionary writing system, which is the lexicographic methodology employed for designing DOPU, goes one step further by automatically extracting word sketch output, namely, collocates, grammatical relations, examples, LCM,[230] together with information obtained with post-processing, like suggestions for labels, and importing them into DWS.

The developers of this methodology (cf. Gantar et al., 2016) have pointed out that one of the advantages of this procedure over analysis of word sketches in the corpus tool is that relevant information necessary for sense disambiguation is already organised in different fields of the database. Such an automation of the process of data analysis streamlines the lexicographic work, saving lexicographers a great deal of time.

In this thesis, this methodology was experimented for designing DOPU. In Chapter 9, I evaluated the level of self-sufficiency of the extracted data in the dictionary writing system for the performance of data analysis for dictionary entry writing. As will be shown in section 11.2 below, some examination of the corpus in the corpus tool was still needed.

Given that, in this thesis the point of departure for identification and distinction of words senses was iLex.

11.1.2 Order of senses

As a corpus-driven dictionary project, this design must apply a frequency-based *order of senses*, not a historical one – like in traditional unabridged dictionaries - or a

[230] Longest-commonest match. See Chapter 5.

logical one (in terms of core senses) – as sometimes used in learners' dictionaries (cf. Atkins, 2008 [1992/3], p. 41; Lew, 2013; Svénsen, 2009, pp. 363-364).

In order to account for sense ordering according to areas of knowledge, which is one of the users' customization features that will be available in DOPU, the plan is to add clickable icons for the great areas of knowledge above the menu area.[231] This way, sense order would be re-sorted – or even a sense would be added or omitted - according to the chosen area. By default, i.e. without any selection of domain, all senses will be displayed and whole-corpus frequency should be followed for senses ordering.

11.2 Data type

Data type can be of four different natures: *internal* to the headword, *external* to the headword, *etymological* and *informative* (Atkins, 2008 [1992/3]).

11.2.1 Data internal to the headword

Lexically relevant data *internal* to the headword refer to the word or the word sense itself, so they should be dealt with in a twofold manner: form and sense. The following table provides an overview of the main issues related to this type of data, adapted to the online nature of DOPU.

[231] As presented in Chapter 10, DOPU adopts a polysemic solution for the headword list, resulting in many long entries. The decision was to use a menu with POS and mini-definition for each entry sense.

Table 11.1 Lexically relevant data internal to the headword (Source: Atkins, 2008 [1992/3])

<table>
<tr><td colspan="2">LEXICALLY RELEVANT DATA INTERNAL TO THE HEADWORD</td></tr>
<tr><td>FORM</td><td>CONTENT</td></tr>
<tr><td>1. Morphological information
* base form
* inflected form</td><td>1. Dictionary sense
* description of the meaning of each lexical unit</td></tr>
<tr><td>2. Orthographic information
* standard and variant spelling
* before and after AOP
* syllabification</td><td>1.1 Meaning types
* denotation and connotation, reference, literal and figurative meaning, cognitive and affective meaning</td></tr>
<tr><td>3. Phonetic information
* IPA
* orthoepic system
* audio with oral pronunciation</td><td rowspan="2">1.2 Meaning explanations
* near-synonyms in the definition
* glosses in the definition
* commentaries
* examples of usage
* formulae
* reference to extra-textual lists of lexical sets (hyperlinking within and outside the dictionary)
* notes
* cross-references drawing users’ attention to related but contrasting entries
* definitions[232]
+ lexicographic definitions: a) traditional format (Aristotelian structure: genus proximum and differentia specifica); b) the COBUILD format (full-sentence definitions); description of the function of the lemma; definitions and synonyms in combination; ‘When’ definitions.
+ defining vocabulary: technical terms; everyday, familiar words; controlled[233] defining vocabulary.
+ graphic illustrations.</td></tr>
<tr><td>4. Lexical form
* single word
* multiword expression
* hyphenated word</td></tr>
</table>

[232] For detailed instructions on how to write definitions, I will resort to Atkins and Rundell (2008) and Svénsen (2009).

[233] It refers to the use of a restricted number of different words for writing the definitions. While it has been adopted by all the renowned English dictionaries for language learners, this practice has not been used for the few Portuguese language dictionaries for foreign learners currently available. For experimental studies on the compilation of a defining vocabulary for a project developing an online dictionary of Portuguese for intermediate-level learners, see Kuhn, Finatto, and Evers (2011), Kuhn and Finatto (2011), and Finatto, Evers, Pasqualini, Kuhn, and Pereira (2014).

11.2.1.1 Form

11.2.1.1.1 Morphological information

In DOPU design, citation forms were used as headwords, namely, infinitive forms for verbs and singular and masculine forms for nouns and adjectives.

In DOPU, nouns and verbs inflections will be displayed after clicking on an icon. The database underpinning VOC, VOP (Orthographic Vocabulary of Portuguese – Portugal), and the dictionaries in the *Portal da Língua Portuguesa*, will be linked to DOPU. This relational database, called MorDeBe (Ferreira et al., 2008), has morphological (verb conjugations and inflection of number and gender for nouns and adjectives) and lexical (variant, loan words, toponymy, deverbal nouns) information of over 300,000 lemmas of different varieties of Portuguese.

11.2.1.1.2 Orthographic information

As informed in the previous chapter, the distinction between standard and variant spellings was not used in DOPU design. Every variant – either between countries (e.g., *econômico* (BP), *económico*(EP)) or within countries (e.g. *característica*/*caraterística* (EP)) – has headword status.

It should be noted that some country-specific variation was introduced by the Orthographic Reform of 1990 (see Chapter 6). There should be a note in this kind of headwords informing that both spellings are officially accepted.

Another important feature of DOPU concerning the spelling reform is that old spellings can still be searched. Users will be taken, however, directly to the headword written with the new spelling, and a note will be provided informing that the searched form is no longer valid. Once again, MorDeBe will be used for linking old orthography to the new one, as this information is available in the database.

Syllabification will be another optional feature, displayed after a click on a specific icon. This information is also available in MorDeBe and will be simply accessed from there.

11.2.1.1.3 Phonetic information

A clickable icon with oral pronunciation of the headword will be available in DOPU. IPA transcription and orthoepic information require development of such material, meaning that, for the moment, they shall not be included.

11.2.1.1.4 Lexical form

As explained in the previous chapter, the headword list in DOPU will comprise single words and multi-word expressions (MWE). Selection of MWE candidates for headword status is not done automatically, though. Based on analysis of single items, lexicographers will point out potentials to be further decided by the editorial team. For example, the MWE *devido a* ('due to') is suggested here as a candidate for headword status. It has a non-compositional meaning and a typical function in sentences, i.e. it functions as a (coordinating or subordinating) connector.

Table 11.2 shows some examples of the different senses of *devido a*. Basically, *devido a* indicates that the idea expressed by the clause in which it appears has a relation of X with the main clause or the other sentence.

It should be mentioned that these are simply concordance lines selected from CoPEP, not examples for the database or the dictionary. This is why some are (slightly shortened) excerpts.

Table 11.2 Different senses of *devido a* in CoPEP

Senses of *devido a*	Examples
Justification	todos os dados foram mantidos, em parte **devido ao** pequeno tamanho da amostra
Reason for	doentes transplantados hepáticos **devido ao** vírus da hepatite C
Cause	Os doentes [com] (HIV) têm alguns riscos acrescidos na endoscopia, como a hipóxia **devido à** infeção oportunística por Pneumocystis carinii

11.2.1.2 Content

11.2.1.2.1 Dictionary sense

As mentioned earlier, the point of departure for word sense disambiguation was the entry in iLex. Automatically extracted data was imported into predetermined database entry fields, namely, headword, frequency, POS, meaning (*glava*[234]) and sense (*pomen*); sense is then composed of a syntactic group (*skladenjske skupine*), which in turn comprises grammatical relations (*struktura*) and collocations (*kolokacije*), with three examples (*zlegdi*) per collocate.

[234] As explained in Chapter 9, I am using the Slovene iLex XML schema.

The headword *candidato* (‘candidate’, noun), ranked among the top 2000 lemmas of CoPEP (rank position: 1056), was used to succinctly demonstrate the process of senses identification and distinction.

Initially, I looked at the collocations under each grammatical relation. It should be remembered that in the automatic extraction procedure for CoPEP, collocates were first clustered, then extracted in groups (of 1 item or multiple items) (see Chapter 9 for more). In Figure 11-1, the grammatical relation displayed is noun (*candidato*) as subject of verb; or roughly speaking, “what candidates do”. Collocates such as *candidato obtém*, *candidato concorre*, among others, can be seen between square brackets.

Figure 11-1Partial view of the entry document for *candidato* in iLex

Another gramrel is *candidato* + *a* + determiner+noun , as seen in Figure 11-1. Just by reading the collocates, without checking the examples, it is possible to begin to distinguish senses. At least four senses can be spotted right away. Four collocates indicating each one of the senses are highlighted in Figure 11-2.

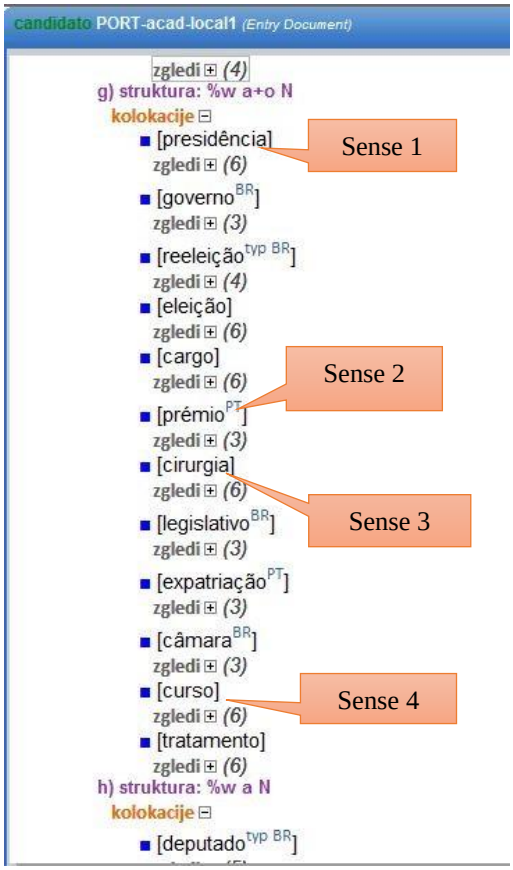


Figure 11-2 Collocations under gramrel *candidato a*+det Noun in CoPEP

By expanding the examples, it is possible to confirm the distribution of senses. The next step is to sort collocates, grouping them according to senses. Figure 11-3 shows examples. It should be noted that fine-grained sense differentiation was not performed at this point.

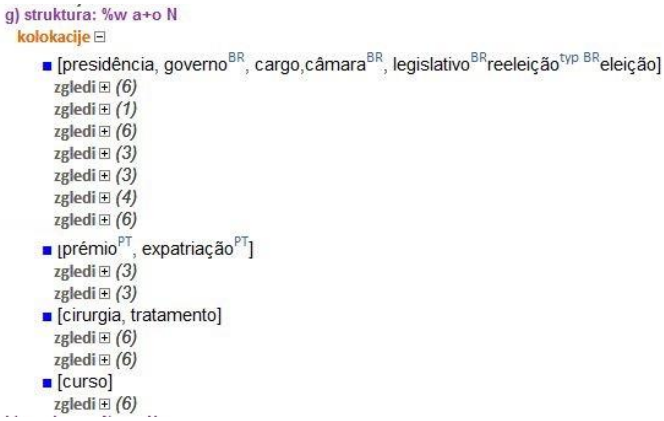


Figure 11-3 Collocates grouped by sense

Semantic analysis of all four senses reveals the existence of one aspect that is common to all four senses, that of evaluation/assessment of candidates. Thus, it could be possible to give one general sense (1.*pomen*), and determine the other four as subsenses (1.1, 1.2, etc). Attention should be drawn to indicators, that work as a mini-definitions to be used in the menu of the entry in DOPU. Figure 11-4 shows a proposal of candidate sense differentiation in the database entry.

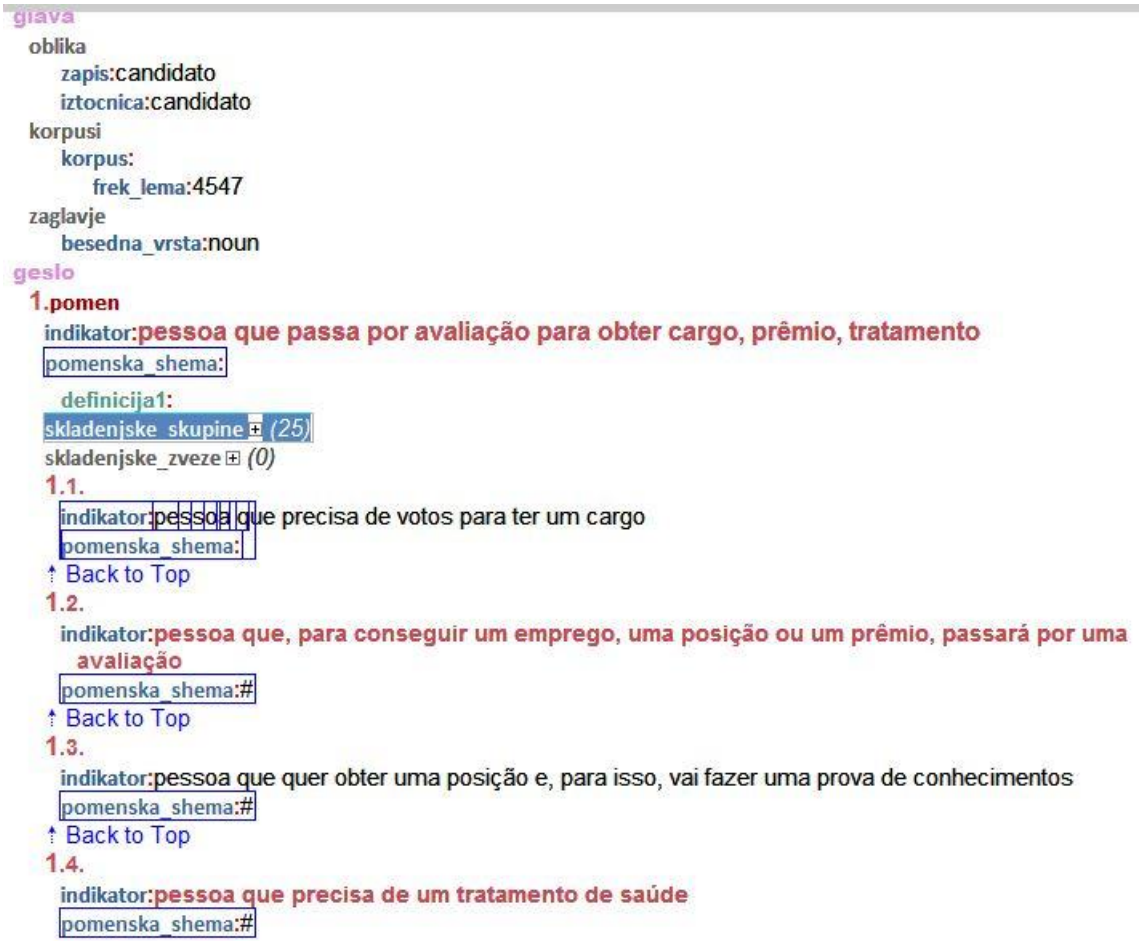


Figure 11-4 Sense and subsenses of *candidato* in iLex entry

A CQL search, in the Sketch Engine, for the structure *candidato* + *a* + infinitive, which was not included in the sketch grammar, revealed another sense, namely, a person that, due to certain characteristics or determined habits, has higher chances of something happening to them. This sense, which would be sense 2 in the database entry, can be seen in this example from CoPEP:

Dentre estas, estão os jovens adultos considerados ***candidatos a sofrerem*** *danos decorrentes da radiação, na medida em que estão frequentemente expostos em atividades de esporte e lazer.*

The next step is to perform a fine-grained analysis in iLex, consisting of the elimination of wrong collocations[235] and bad examples, validation of good collocates and examples, sorting of misplaced collocates, sub-division of senses, and addition of indicators.

11.2.1.2.2 Meaning explanations

According to Atkins (2008 [1992/3], p. 44), meaning explanations designate "all the information in a dictionary entry that the lexicographer employs in order to transmit the meaning of an item to the dictionary user".

a) Synonyms and near-synonyms in the definition

   Synomyms and near-synomyms shall be displayed separately from the definition in a special area that can be easily seen.

b) Glosses in the definition

   Glosses might be employed when rare words and words related to history, culture, or discipline are used. A non-intrusive solution for using glosses without making the definition text too dense is to adopt a mouse-hover feature in DOPU. Hence, when the cursor hovers over a word that is glossed, the gloss is shown in a caption above the word. The online dictionary *Infopédia* uses a similar approach. However, it gives the definition of every word, instead of a gloss. Figure 11-5 shows an example.

[235] As explained in Chapter 7, "wrong" indicates that the collocation does not reflect the grammatical relation under analysis.

Figure 11-5 Mouse-hover feature in *Infopédia*

c) Commentaries

Will be given when lexicographers understand it is necessary to provide the users with a longer commentary so that meaning is clarified.

d) Examples of usage

Examples are a fundamental feature of DOPU, accordingly, a GDEX configuration was developed (see Chapter 8 for more) in order to help lexicographers with the very demanding task of example selection.

e) Formulae

Considering that CoPEP comprises texts of different areas of knowledge, formulae will be used to support definitions of, for instance, chemical substances. This is one of the situations when experts in the area might be summoned to contribute with specialised knowledge.

f) Reference to extra-textual lists of lexical sets (hyperlinking within and outside the dictionary)

It seems apparent that hyperlinking to different glossaries of terms, for instance, could be very useful for DOPU users. Additionally, certain concrete words could have hyperlinks to a bank of image (something like Google Images) to complement the definition of certain words.

In terms of hyperlinking within the dictionary, it seems useful to make words clickable so that users are taken to their entries.

g) Notes

These are highly relevant assets in a dictionary as they almost play the role of a teacher providing additional information to the user in a simple and condensed

way. For academic language, it is only natural to make use of notes in certain entries or sub-entries.

h) Definitions

There are several ways of writing lexicographic definitions, namely, the traditional format (Aristotelian structure: *genus proximum* and *differentia specifica*); the COBUILD format (full-sentence definitions); description of the function of the lemma; definitions and synonyms in combination; and 'When' definitions.

While for certain lemmas some explanations are better given with one certain definition style rather than another, it is important to maintain consistency throughout the dictionary with the applied variant. A guideline will have to be prepared by the editorial team before implementation of the project for making DOPU.

i) Defining vocabulary

A brief overview of the use of a controlled defining vocabulary was given in Chapter 4. The decision made for DOPU design was that it would not be used. Lexicographers will be instructed to avoid rare words, unless the lexical unit being described requires the use of specific terms or technical words.

11.2.2 Data external to the headword

For Atkins (2008 [1992/3]), lexically relevant data *external* to the headword involve the presence in the dictionary of facts relating to the headword's relationship with other words in the language. These data comprise four classes: relational, paradigmatic, usage, and syntagmatic information.

Decisions about the inclusion of morphological facts like derivation (run-ons, subentries) and cross-references to other entries in the dictionary fall into the *relational information* class. Yet facts concerning verb conjugation, parts of speech, synonyms and antonyms refer to *paradigmatic* information.

Syntagmatic information entails decision-making with regards to (Atkins, 2008 [1992/3], pp. 47-48):

- Complementation: designates “the range of syntagmatic environments in which a word’s full semantic potential may be expressed”. The question is: which constructions must be recorded as part of the “active scope” of the headword? Atkins states that this issue relates to the linguistic issue of arguments versus adjuncts;
- Collocation: designates “the significant co-occurrence of two or more words (either because such grouping happens more frequently than the statistical norm, or because it represents a semantically opaque lexical item).” The author’s concern refers to dictionary designers’ (editors’) difficulty involved in establishing classification criteria for collocation so that there is consistency in the treatment of the language in the dictionary.

For Atkins, *usage* information refers to labelling language facts. She presents some possibilities like register (formal, informal), currency (obsolete, old-fashioned), style (poetic, technical), pragmatics (expressing pleasure), status (dialect, slang), and field (architecture, music). Corpus analysis will provide information on which components to include in the entries. For now, I will assume that, given the characteristics of the corpus as representative of current academic language, some language realities will not be frequent enough in order to make it into the dictionary (such as register and currency, for instance), while others, like field - or in DOPU’s case, area of knowledge - shall be a label.

11.2.2.1 Etymology

In relation to DOPU design, the *etymology* of headwords was not included. As shown in dictionary user research, users tend to not consult this element in entries. Nevertheless, it could be possible to provide hyperlinks to reliable online etymological dictionaries.

11.2.2.2 Informative data

Informative data relate to “useful comments that the compiler may make in order to clarify a particular entry” (2008 [1992/3], p. 43). Atkins and Rundell (2008, p. 233) add that these data in the entry aim “to tell their [the dictionaries’] users what they need to know, even when this will not fit the model of the traditional dictionary entry”. Thus, they might refer to notes specifically concerning the subject of the entry (for instance, in

DOPU's case, a note explaining the use of the definite article in front of the names of the disciplines) or the headword proper, such as the "Get it right" boxes in Macmillan and the different *notes*[236] employed by the *Oxford Learners' Dictionary of Academic English*.

11.3 Additional information

An important issue that is on the borderline of macro and microstructural decisions is the frequency marker. Some monolingual English dictionaries for learners have adopted this system, each using a different set of symbols.[237] Ideally, frequency of occurrence should be designated to the lexical unit (dictionary sense) level. Unfortunately, automation has not reached this level of fine-grained analysis yet, so information on frequency can only be differentiated in terms of homographs.

The ranked CoPEP lemma list can be seen in Appendix D (CD-Rom). It contains the top 2,000 lemmas, separated into two groups: top 1,000 lemmas and top 1,000 to 2,000 lemmas.

[236] "Language bank" is an example of note which indicates the textual function of certain words, like "according to – reporting someone's opinion". For a detailed description of the usage notes of the *Oxford Learners' Dictionary of Academic English*, see Kuhn (2015).

[237] Learner's dictionaries have adopted different systems to pass that information in an easy, explicit way. As an example, the Macmillan English Dictionary for Advanced Learners uses a red-star system, in which "three-star words are the most common 2,500 words in the language. Two-star words are the next most common, and one-star words are the next most common 2,500" (http://www.macmillandictionary.com/learn/red-words.html). *Longman* uses boxes with information about frequency in written and spoken modes. For instance, W3 indicates that the headword is one of the 3,000 most frequent written words.

Chapter 12 Discussions

This chapter begins with a review of the semi-automated approach employed in DOPU design. More specifically, it discusses the development of the resources required for automatic extraction, namely CoPEP (12.1.1), sketch grammar (12.1.2), and GDEX (12.1.3). It finishes with an analysis of the advantages and limitations of automatic extraction of data and import into a dictionary writing system in the context of creating a dictionary of academic Portuguese.

In 12.2 I review DOPU, highlighting its advantages over other dictionaries of Portuguese regarding tertiary students' needs (12.2.1), suggesting enhancements (12.2.2), and pointing out DOPU's potential publication.

I then move on to the presentation of the contributions of my PhD research (12.3), finishing the chapter with some pointers for future work (12.4).

12.1 Review of the semi-automated approach

As Gantar et al. (2016, pp. 218-219) report, comparison of the manual approach (analysing and selecting relevant data in the corpus tool) and the semi-automated approach showed that the latter is more effective and time efficient, and streamlines the lexicographical process without reducing the quality of the information provided in the dictionary.

In the same vein, a general balance of the experiment with the semi-automated approach for designing DOPU has also proved to be undoubtedly positive. This PhD research has demonstrated that implementation of this approach for Portuguese lexicography is indeed possible.

Given the novelty of this method in lexicographical projects for dictionaries of the Portuguese language, it is not surprising that some challenges were encountered along the way. In the next sections, I will discuss some of the advantages and shortcomings of this approach, focussing on the new resources and tools that had to be devised, as well as the modifications in the procedure of extraction.

12.1.1 CoPEP

One of the major advantages of having built CoPEP as a carefully-designed corpus was the level of minutia that was possible to achieve when creating the design. Thus, entry compilation, for instance, could go much further towards sense differentiation, usage explanation, and the selection of discipline-derived examples than other dictionaries of Portuguese that were used as a source of comparison.

There were limitations, however. One of the most relevant for designing DOPU was the multi-variety, multi-norm nature of CoPEP. The effects of the recent spelling reform on Portuguese NLP have only just begun to be addressed by computational linguists. Portuguese resource developers remain divided between a variety-specific approach and a multi-variety perspective, with the former being highly favoured. In consequence, there were no easily applicable solutions to deal with CoPEP.

Given this situation, post-processing of CoPEP and the Freeling tagger were workaround measures to allow development of my research in a rigorous and (as much as possible) accurate manner, bearing in mind the conditions of production of this thesis.

CoPEP_AO90 was used for developing the final versions of Sketch Grammar and GDEX and for the second procedure of automatic extraction of data from the corpus and import into iLex, meaning that DOPU was designed, i.e. macrostructure was defined and pilot-entries were compiled, using CoPEP_AO90 as a source of lexical information. This was only possible due to the processing of Freeling as well.

12.1.2 Sketch grammar[238]

The Sketch Grammar for Academic Portuguese, developed for exploring grammar and lexis of Portuguese in the CoPEP corpus, has had implications not only for my work on the dictionary of Portuguese for university students, but also for Portuguese corpora in general. A comparison with the default sketch grammar available for Freeling-tagged corpora of Portuguese reveals that AcadPortSkG comprises a larger number of grammatical relations for nouns, verbs and adjectives, and brand-new rules for adverbs, thus broadening up word class coverage. In addition, the queries of existing

[238] This section was published in Kuhn and Kosem (2016).

sketch grammars, which were used in developing AcadPortSkG, were adapted and now yield better results. Lastly, AcadPortSkG contains queries which were carefully devised in a way to overcome detected annotation errors, making it more accurate. It should be mentioned that the latest version (see 7.3.3) has enhanced this grammar even further.

This new grammar, with broader coverage and more complex gramrels, yields very rich results –most of the time it produces more data than can be handled by a human, which is, in fact, typical of sketch grammars for automatic extraction of lexical data (Kosem et al., 2013). The breadth and depth of the advantages of this sketch grammar can be seen in the results of the process of extraction of data from CoPEP and import into iLex shown in Chapter 9.

There is no doubt that AcadPortSkG can also be used with any other corpus of Portuguese tagged with Freeling v.3. Such corpora will benefit from my sketch grammar on two levels. In terms of their utilization, the users of the corpora will be able to conduct a more thorough and reliable lexical analysis due to a greater number of grammatical relations and their (improved) accuracy. Concerning the development of tools for Portuguese, the very process of sketch grammar development has revealed problems with corpus annotation that can be used to improve the tagger and inform other resource developers of potentially problematic areas. One such improvement has already been implemented by the Sketch Engine team. My identification of problems with tokenization and lemmatization of verbs with *-se*, together with a detailed explanation of that particle function and a thorough description of faulty annotations, led to the reconfiguration of the Portuguese pipeline in the system.

There is still plenty of room for improvement, and I provide a few suggestions here: a) use of macros in the language m4. Macros are used to avoid repetition of recurring patterns. For instance, a macro "adjective" can be defined that includes adjectives and participle forms, thus "A.*|V.P*" does not have to be written every time adjectives are included in queries; b) corpus annotation. Reporting errors in tagging have led the Sketch Engine team to make some significant modifications in the pipeline for Portuguese corpora. Further collaboration can help overcome existing shortcomings that could not be addressed at this time; c) enrichment of the sketch grammar by devising queries for grammatical relations that are currently not covered.

Although this sketch grammar can be further improved, the current version already yields very good results, for both academic and general Portuguese. Thus, I decided to make it available in the Sketch Engine for researchers using/making Freeling-tagged corpora of Portuguese.

12.1.3 GDEX

One of the most significant findings resulting from the whole process of GDEX configuration for academic Portuguese is that values of certain features of CoPEP obtained from statistical analysis are different than those in a 3.9 billion-word reference corpus, namely, Portuguese Web 2011. Among them, I highlight here the considerable difference in sentence initial tags, as is shown in Table 12.1 below. Tags are ranked in order of frequency.

Table 12.1 Rank of sentence initial tags in CoPEP and Portuguese Web 2011

CORPUS RANK	COPEP 40 MILLION WORDS	PORTUGUESE WEB 2011 (PTTENTEN, FREELING V3) 3.9 BILLION WORDS
1ST	Verbs: 181,197 (3,725.17 per million)	Nouns: 42,800,096 (9,250.90 per million)
2ND	Prepositions: 167,601 (3,445.65 per million)	Prepositions: 'empty result' (?) Verbs: 26,855,800 (5,804.70 per million)
3RD	Adverbs: 144,594 (2,972.66 per million)	Adverbs: 17,126,081 (3,701.70 per million)
4TH	Nouns: 143,198 (2,943.96 per million)	Prepositions: 26,249,572 (5,673.60 per million)
5TH	Conjunctions: 77,235 (1,587.85 per million)	Conjunctions: 10,800,586 (2,334.50 per million)
6TH	Pronouns: 44,408 (912.97 per million)	Pronouns: 9,639,404 (2,083.50 per million)
7TH	Adjectives: 17,972 (369.48 per million)	Adjectives: 1,395,167 (301.60 per million)

This particular preference for sentence initial word categories might not only be indicative of specialization of language register, that is, academic language, but also

points out differences between languages. In the case of Estonian (Koppel, 2017), analysis of the training dataset containing good examples indicated nouns as top-frequent sentence initial tags (similarly to general Portuguese), followed by adjectives and pronouns in almost the same proportion (which do not make the top-five rank in both corpora of Portuguese), with verbs ranking the lowest, while in academic Portuguese it takes first and second position in general Portuguese.

Another interesting finding is that average sentence length of academic Portuguese is considerably longer. This seems to corroborate the seminal work on the characteristics of academic texts written in Portuguese (mostly European) carried out by Bennett (2008, 2010a).[239] Manual examination of a corpus of around 400 texts indicated that Portuguese academic writing tends to be verbose, with long, embedded sentences, contrary to Academic English, which is the model widely used for Portuguese academic writing instruction in Brazilian and Portuguese higher education institutions. This is highly significant for DOPU users as they will be offered evidence of real usage of academic Portuguese as a qualified, theoretically-sound alternative for those misguiding examples translated from English and provided in class.

Although examination of human-judged examples is ideal for the provision of accurate model data, the fact that I used the statistic-based description of lexical and syntactic behaviour of the language used in the corpus foregrounding the dictionary had a major influence on the quality of GDEX output. It can be concluded that another important aspect to be carefully considered in the development of GDEX configurations is language register, at least when it comes to Portuguese.

In the specific context of classifier development for the corpus of academic Portuguese, a crucial finding concerns the fact that the top collocate examples yielded in the GDEX editor only partially cover different meaning patterns (cf. Kosem, 2010, pp. 211-223), irrespective of the areas of knowledge to which the examples belong. One illustrative example concerns the highly frequent adverb *através* (rank 131; 536.00 occurrences per million), whose most frequent grammatical relation (29.98%) in CoPEP is verb modified by adverb (V mod por através_Adv).

[239] See section 2.4.2 for a detailed review of these work.

Performance of a word sketch for the headword and manual analysis of concordance lines of collocates with salience higher than 7.0 indicated different syntactic patterns.[240] These are three of the most frequent patterns identified:

1. noun phrase+ passive verb+ *através de* + noun phrase
2. *ser*+ *através de* + noun phrase + *que* ('be+ through +NP+ that')
3. noun phrase + active verb+ *através de* + noun phrase

Among a series of different meanings found in all three patterns, one is especially highlighted here, deriving from syntactic pattern one. The collocation under examination was *transmitir* ('to transmit') *através de*:

noun phrase 1 [disease, bacterium, virus, pathogen] +*transmitir* passive+ *através de* + noun phrase 2 [animal, saliva, tear, (vegetable) bud].

In this pattern, *através de* indicates that elements in NP2 have the pathogenic agent within and function as its carrier, transferring it to a living organism, which is usually implied from the context, without being named in the sentence.

This meaning is used in Health Sciences and in Agricultural Sciences. However, it cannot be identified in the top three examples (as planned for the automatic extraction procedure) yielded by GDEX. Such limitations with GDEX sorting has also been noticed by Kosem (2010) with a corpus of academic English. These results raise a question as to the extent to which GDEX is useful in extraction of examples of corpora of academic language, given that they should account for the diversity of discipline-geared linguistic behaviours.

These findings also indicate that more investigation is needed to evaluate the influence of areas of knowledge on the efficiency of classifiers. It seems plausible to hypothesize that different domains could require specific values or even new classifiers.

The next step in the development of GDEX for Academic Portuguese involves integration of statistics obtained from human-evaluated examples and comparison of the

[240] At this point attention should be drawn to the fact that através occurs 99.99% of the time as part of the multi-word lexical unit *através de* ('through, by, throughout', among others) in CoPEP. In fact, occurrences without "de" actually derived from tokenization problems and errors. Thus, pattern analysis focused on the elements to the left and right of *através de*.

results of the two configurations. The analysis might contribute to further fine-tuning of values, or even reveal potential for new classifiers as the Estonian case shows.

12.1.4 Automatic extraction of data procedure

The experiment of automatic extraction of data and import into DWS was performed on pre-final versions of required resources, as mentioned in Part II. Thus, this experiment contributed to indicating those parts requiring improvement, which was implemented in the latest versions of CoPEP, sketch grammar and GDEX configuration, each one thoroughly explained in Chapters 6 to 8. Moreover, although the development of the extraction process should continue, information obtained in this extraction was deemed satisfactory for the purposes of this thesis and was used for the design of DOPU (Chapters 10 to 12).

Continued development of the procedure together with its fundamental elements, looking for quality enhancements not only of the process of automatic data extraction, but also the final product, i.e. the dictionary, would most certainly be ideal. However, this ambitious enterprise demands a dedicated team of lexicographers, linguists and computational linguists, as well as proper allocation of time and, very importantly, funding.

As the experiment carried out in this PhD research demonstrated the potential for adopting this approach for Portuguese lexicography, some initial plans for a final extraction have already been made. Whether or not the project is, in fact, implemented, I will carry out another extraction, especially due to the possibility of making these automatically extracted data available online for users (as mentioned earlier).

A very important lesson learnt is that the quality of corpus annotation is crucial for automatic corpus analysis. Since the GDEX algorithm is based on word sketches, problems with wrong tagging severely interfere with word sketch results, consequently generating non-representative examples. Additionally, Portuguese is a language that allows some positional variance, resulting, for example, in many instances of subject and object canonical position (subject + object) inversion and embedding of clauses between subject and object. Thus, it has been shown that word sketches based on tag positions may be faulty, especially for certain grammatical relations like subject and

verb. The lesson to be taken here is that corpus annotation is crucial for yielding good examples.

One additional issue that deserves more investigation in the future of data extraction for DOPU concerns the distribution of occurrences across different texts. More specifically, is it possible to include this type of information in the extraction script?

The case below illustrates how the procedure currently functions. In iLex, the collocation *carreiras habitacionais* occurs only in EP. Verification of the areas of knowledge in the automatically extracted examples revealed Applied Social Sciences (AP). However, as there are only three examples per collocate, I wanted to confirm if AP was the only area of knowledge where this collocation occurs. I then went to my corpus in the SkE and performed a text type frequency filtering.

Figure 12-1 Concordance lines for the collocation *carreiras habitacionais*, while Figure 12-2. indicates that they all occurred in one text only.

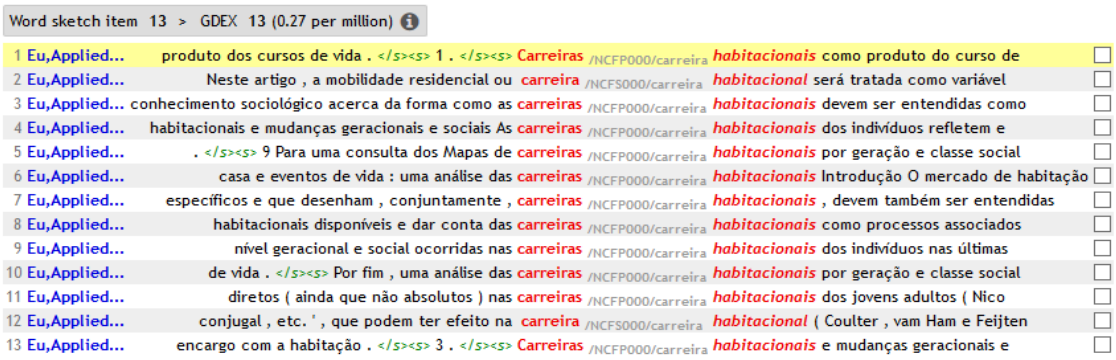


Figure 12-1 Concordance lines for the collocation *carreiras habitacionais*

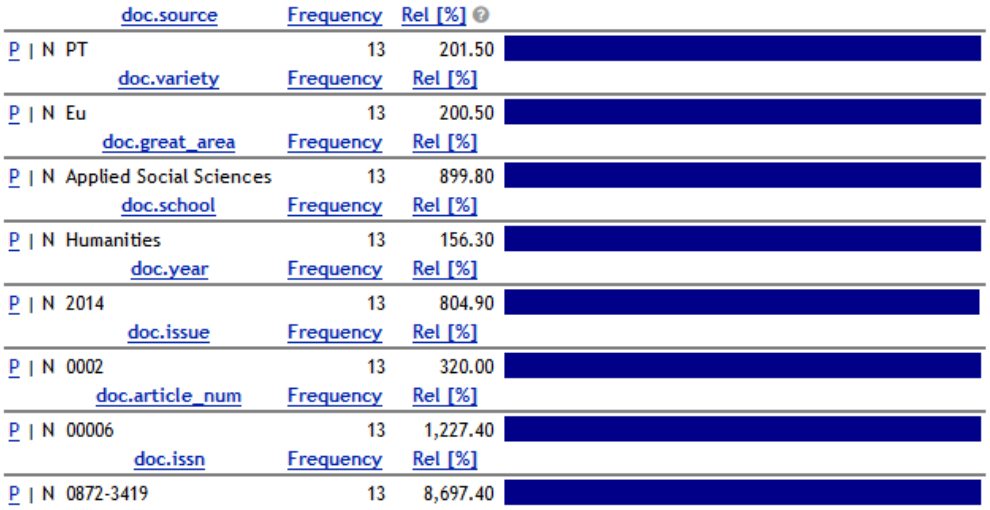


Figure 12-2 Collocations in one text only.

One of the most significant findings from the development of the GDEX configuration for the corpus of academic Portuguese is that example sorting based on the whole corpus can be biased towards one area of knowledge. This is a shortcoming as it is necessary to go back to the corpus to verify whether such collocation happens in other areas of knowledge as well.

12.2 Review of DOPU

To review DOPU means to evaluate its potentialities given what has been achieved by developing its design.

12.2.1 Advantages of DOPU

The advantages of DOPU are of at least two natures. On the one hand, it refers to the great use of its format, namely, online. On the other, it concerns the compliance with its target-users' needs.

As demonstrated throughout this thesis, there are several possibilities to implement state-of-the-art digital techniques to make DOPU simple for the user but still rich in content. A web designer and a programmer will obviously be responsible for such developments.

Undoubtedly, the kind of lexical information that will be displayed in DOPU is much richer than the monolingual dictionaries of Portuguese currently available. One example is the headword *candidato* (see Chapter 11), which contains many more lexical units than such dictionaries.

Additionally, one of the major assets and a unique feature of DOPU is the presentation of collocations. Even without further fine-grained analysis, users can already benefit greatly from the examples provided for each collocation, as they help identify syntactic and semantic patterns.

12.2.2 Suggestions for enhancement

Undoubtedly, in terms of content, improvement of DOPU depends on the enhancement of resources and tools used for building it. In this sense, as these lexicographic requirements are constantly being developed, it is expected that DOPU will, as time goes by, present improved entries.

As to users' reception of DOPU, there are a number of aspects to be observed, such as usability (do users find it simple to use?); content relevance (do users find what is available useful?); coverage (does DOPU provide everything users need?), among others. There are ways to tackle these issues: prompting the user to provide answers; providing a method of contact for pro-active users to come forward and make comments; or by automatically analysing individual users' activities in the dictionary:

- **User research**: However, instead of adopting the traditional method of collecting answers from a certain sample, a possibility is to present a quick online quiz that would pop up in a smaller window when users access the page or maybe when they click on the close icon. Mostly, they would be asked about their satisfaction with the dictionary (apps have been using this strategy), with very simple yes/no questions and a Likert scale.
- **Contact**: besides the traditional display of an email address for contact, a "send a message" area should be available. The main difference is that in such an area, anonymity would be guaranteed. Additionally, links to social media like Facebook and Twitter should offer users additional means of expressing opinions.

- **Log-files**: this practice has been used for the observation of dictionary users' look-up habits for quite some time. The downside is that it requires a login. It has been reported that users tend to reject applications which require logins. Nevertheless, one alternative to make it simpler for users to log in is to allow direct logging in via Facebook. This is a practice very much used nowadays for accessing several applications, and in this sense, digital users are familiar with the process.

12.2.3 Potential publication

With the advances of this PhD research, it is possible to make DOPU immediately available online, at least as a work-in-progress resource. Dictionary entry output containing automatically extracted collocations and examples can be published and students can already have access to it. Certainly, a note should be provided indicating the unreviewed status of the entries. Nevertheless, it is believed untreated data is better than no data at all.

Ideally, DOPU would be already connected with MorDeBe database by the time DOPU is put online, thus additional information such as inflections and variants could be displayed without previous review.

Although the form of presentation of the stages here implies the adoption of a traditional routine for dictionary-making that broadly revolves around the idea of three consecutive phases, namely, planning – writing – producing (cf. Landau 1984: 227 as cited in Klosa, 2013, p.517), it cannot be forgotten that the advent of electronic lexicography (see Chapter 3) has led to significant changes in the lexicographical process.

In this vein, Klosa has specifically proposed a computer lexicographical process for online dictionaries under construction (Figure 12-3). According to the author, unlike a lexicographical process for printed dictionaries, a computer-lexicographical process for online dictionaries under construction is characterized by the merging of different phases. This subverts the chronological sequence of the stages and allows for independent lexicographic production. Hence, a certain module of a dictionary might be

published, while other parts are still under development. It is no longer necessary to wait for the dictionary project to finish to make the product public.

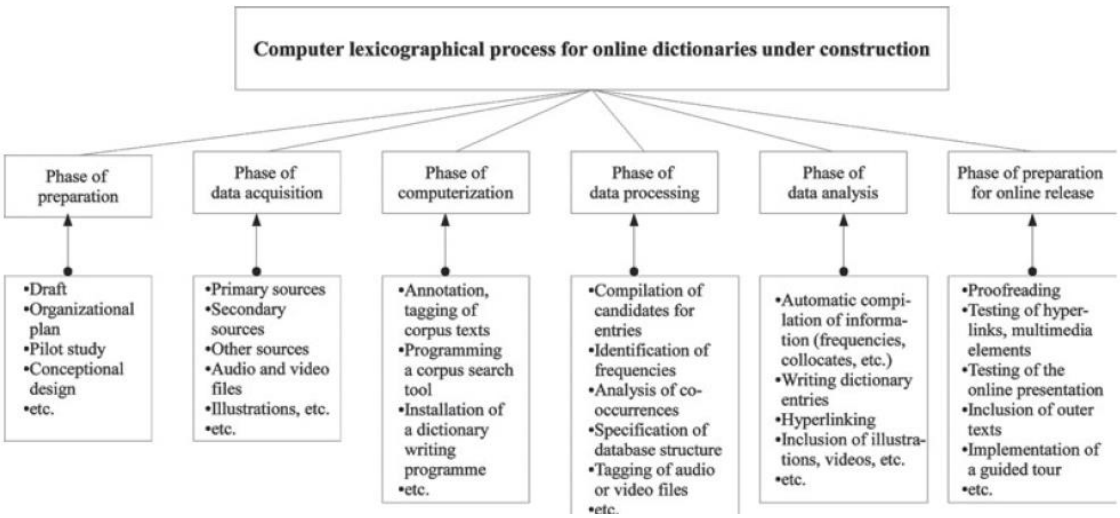


Figure 12-3 Reproduction of the visualization of a computer- lexicographical process for a corpus-based online dictionary under construction proposed by Klosa (2013, p. 520)

At this point it is important to remember that although this thesis is about the design of DOPU and not the dictionary proper, the lexicographical process as explained by Klosa (2013) was followed closely. It could not be different given that DOPU is a digitally-born online dictionary. The main difference was, evidently, the scale of the actions performed in each phase.

12.3 Contributions of this research

The undertaking of this PhD research has made a number of significant contributions, especially to four areas of knowledge: lexicography, linguistics, pedagogy, and linguistic policy.

a) Lexicography

In terms of lexicographic products for the Portuguese language, the design developed here has provided the principles and parameters for making a novel, up-to-date dictionary of Portuguese specifically tailored for university students. It is hoped that lexicographers take on this project and implement it.

The employment of state-of-the-art methodology for dictionary-making, namely, the automatic extraction of data from the corpus and import into the dictionary writing system, is in itself one of the major contributions to the area. Due to the undertaking of rigorous experiments, it is now possible to build on the findings of this research to adjust this approach according to Portuguese language needs.

Additional major contributions are the unprecedented resources and tools that I created in order to apply the semi-automated approach, namely:

- *The CoPEP corpus*. It is the first carefully-designed, balanced multivariety corpus of academic texts. It is planned that CoPEP will be publicly available for online consultation and download from the CELGA-ILTEC website as of October 2017.
- *Sketch grammar for academic Portuguese*. It should be noted that this is the first sketch grammar for Freeling-tagged corpora or Portuguese that broadly covers various grammatical relations and is free of malformed queries.
- *GDEX configurations*. An AcadPort-GDEX configuration was created for finding good examples in Portuguese academic texts. This resource was the first ever developed with full attention to a series of classifiers exclusive to the Portuguese language.

Overall, it can be said that the employment of the most advanced technology for dictionary-making equalizes the lexicography of the Portuguese language with other further-developed traditions and internationalizes Brazilian and Portuguese research.

b) Linguistics

The corpus of academic texts will make an enormous contribution to corpus linguistics, descriptive linguistics, contrastive linguistics, academic discourse analysis, to name but a few.

For corpus linguistics, the contribution is twofold. On the one hand, the methodology of compilation could be used as a guideline for future compilation of other corpora, by other researchers, or as a starting point to build on. On the other hand, the area is enriched by the availability of a unique corpus, with particular characteristics–

covering Brazilian and Portuguese varieties in a balanced way, composed of peer-reviewed online journals, representing texts from four different areas of knowledge.

This corpus is a rich, sound source of information for obtaining in-depth knowledge about academic Portuguese. Its synchronic, area-varied characteristics allow linguists to describe authentic, current language use in Brazilian and Portuguese academe.

Brazilian and Portuguese varieties are equally represented in this corpus, thus allowing for data analysis to reveal distinctive characteristics which contribute to the area of contrastive studies.

c) Pedagogy

The primary goal of this PhD research was to propose a design of an auxiliary lexicographic tool for university students engaged in literacy activities in higher education. It is unquestionable that DOPU is also a rich resource for teaching-learning academic Portuguese.

A secondary implication of this PhD research for pedagogy refers to the development of linguistically sound teaching material of academic Portuguese due to language analysis of the corpus compiled here. The possibility of finding structured information of authentic language use, from collocational behaviour to vocabulary tendencies, facilitates the work of teachers.

The development of DOPU and, potentially, other teaching materials, contributes to filling in a significant gap in the research-informed production of tertiary-level educational tools in Portugal, Brazil and other CPLP countries.

d) Linguistic policy

This PhD research aimed to propose a dictionary of academic Portuguese (DOPU) as a pedagogical lexical tool to be used by speakers of Portuguese as a mother tongue and PAL who study at universities all over the world. The very development of this design is in line with the recommendations put forward in PALis, namely, those that refer to the dissemination of Portuguese as a language of science, the need for the development of teaching material for speakers of PAL, the suggestion of an online resource concerning different varieties of the Portuguese language, and a focus on the

production of material for teaching and learning Portuguese for specific purposes. The effective production of DOPU will put those recommendations into practice.

Furthermore, the undertaking of this PhD project in Portugal by a Brazilian researcher, under the supervision of Portuguese linguists, reveals the underlying motto of my research: to strengthen the academic relationship between the two countries and to establish an international team that can take on the lexicographic project to make DOPU.

12.4 Future work

In the case that the lexicographic project for DOPU is taken on by a team of Brazilian and Portuguese researchers – linguists, lexicographers, computational linguists, web designers - there is a series of related studies that could be advanced in addition to the actual make of DOPU. Below, I list some of them:

a) Broadening the national varieties of Portuguese included in the dictionary by working in partnership with researchers from other CPLP member states;

b) Developing an international survey with university students of all eight CPLP countries to obtain a clear, reliable and up-to-date description of the real use of dictionaries of Portuguese. These descriptions can then later be used to implement important modifications in DOPU, thus better fitting users' needs. The outcome can also contribute to the improvement of monolingual dictionaries in general;

c) Increasing the size of the corpus with other genres and documents covering other language varieties;

d) Compiling structured subcorpora of university students – both speakers of Portuguese as a mother tongue and as PAL;

e) Studying different genres and different parts of each genre, to obtain detailed descriptions of the way texts are composed in Portuguese (and in each of its varieties), and special uses of the language in each situation;

f) Developing the Academic Portuguese platform/portal which would host not only the dictionary, but other features like:

- Terminological resources related to academic/scientific knowledge (glossaries, vocabularies, bilingual specialized dictionaries);
- "My vocabulary": a function where the system records all the searches made by the user and keeps them in a special area, which can be easily accessed and reviewed by the user;
- "Vocabulary builder": a function where the system keeps a record of the words that have been searched by the user in order to propose lexical exercises. Some possibilities are flash cards that could appear in a pop-up window every time the user logs in;
- "Writing tutorial": a section with explanations about each kind of academic genre (abstract, article, report, etc.) and descriptions of language use;
- A corpus query tool for users wanting to plunge deeper into language description. The Skell program is an inspiration;
- A contribution area for teachers where teaching materials can be shared/downloaded;
- Lexical gain exercises, to be available online, so that they can be interactively answered by users with downloadable versions for teachers to take to their classes.

Conclusions

Studies have shown that academic language has its own characteristics. The recognition of the importance of teaching/learning academic language has led to investments in study and pedagogical material development by researchers of many languages, including calls for dictionaries of academic language (e.g. Kosem, 2010; Granger and Paquot, 2010a, 2010b; the *Oxford Learners' Dictionary of Academic English*, 2014). Given the dominance of English as an academic lingua franca around the world, it is not surprising that most efforts into academic language research and material development have been made for English. Nonetheless, it should not be forgotten that other languages also have their own academic languages which consequently also need resources such as dictionaries and materials for teaching and learning academic language. As was shown throughout this thesis, this is also the case of Portuguese.

Significant expansion of access to higher education - in Brazil in the last decade; in Portugal in the 1990s – has resulted in a great number of university students studying in Portuguese. In addition, the current phenomenon of internationalization of Portuguese due to an increasing economic interest in Portuguese-speaking countries like Brazil and Angola, has led to major growth in the broad area of Portuguese as an Additional Language and, in particular, to an increased number of speakers of other languages pursuing their studies in universities where Portuguese is the medium of instruction, i.e. in member states of the Community of the Portuguese Language Countries or countries where Portuguese is not the official language. Those students face an even greater literacy challenge: to master academic language skills in an additional language.

As a result, we have also witnessed an increased interest in literacy practices in higher education and the first attempts of describing academic Portuguese. Nevertheless, the area of academic language studies is still rather incipient in Brazil and Portugal, thus research-informed teaching/learning resources for academic literacy in Portuguese are scarce. One pedagogical tool that is missing is a dictionary of academic Portuguese.

My PhD research was thus an effort to contribute to filling in this significant gap by proposing an online corpus-driven design of a dictionary of Portuguese for university students (DOPU).

The key part of my proposal was the adoption of the semi-automated approach to dictionary compilation, as originally proposed by Rundell and Kilgarriff (2011) and first implemented into lexicographic practice by Gantar et al. (2016). In this approach, lexical data (grammatical relations, collocations, examples) are automatically extracted from the corpus according to predetermined criteria, and transferred to the dictionary writing system where lexicographers then analyse, validate and edit the data to shape them into the final entry.

As a method that had never been adopted for lexicographical projects of the Portuguese language, my research aimed at experimenting the approach for the first time. It is not surprising that such an innovative project in the context of Portuguese posed a series of challenges, mostly due to a lack of the prerequisites that are required for implementation of the method. According to Gantar et al. (2016: 220-221), the procedure of automatic data extraction, which uses the Sketch Engine tool (Kilgarriff et al. 2004), is language-independent, although the following requirements have to be met: a relatively extensive corpus, POS-tagged as accurately as possible, a sketch grammar, GDEX configuration(s), and parameter settings for extraction.

In consequence, the core of the design proposal of DOPU was to develop these missing resources and tools. Firstly, I compiled the CoPEP - *Corpus de Português Escrito em Periódicos* ('Corpus of Portuguese from Academic Journals'). Next, I devised a sketch grammar specifically for CoPEP and DOPU conceptualization purposes. Then, I developed a GDEX configuration for academic Portuguese. Finally, I prepared a DOPU-oriented version of the procedure for automatic extraction of data in which important additions such as assignment of language variety labels were introduced.

At the end of this journey, I am very happy to observe that the experiment of automatic extraction of data from CoPEP and import into iLex was successful, proving that this approach is also applicable to Portuguese. As an enthusiast of this method, it is my intention to continue working on ways to improve the procedure. I am certain that

implementation of DOPU can start from the design I proposed in this thesis, and resources, tools and procedures can be improved along the way.

Looking ahead

As has been shown on numerous occasions throughout this thesis, automation has been enabling faster, qualified developments of lexicographic products. It is unquestionable that the future of lexicography will involve deeper connections with NLP and computational advances.

One of the areas in which dictionaries will play an essential role is writing assistant programs. Reading and writing have become routine activities that, in some cases, take up most of a person's time. This is unprecedented in the history of humanity. It is widely known that a significant part of social life has been taking place in spheres dominated by the written language. Distance learning requires well-elaborated text exchanges for common activities like posting questions in forums or working on collaborative projects. Work of the most varied areas and kinds use email as a common means of communication.

Overall, it is apparent that dictionaries will not disappear in the near future; they will assume different formats and functions. And we, lexicographers of the pre-digital natives' era, will have to follow suit.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. doi:10.1016/j.jeap.2013.08.002

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. doi:10.1016/j.esp.2011.08.004

Alle, C. M. de O. (2009). *Estudo exploratório de verbos causais em pediatria: padrões em traduções do português para o inglês*. Retrieved from http://hdl.handle.net/10183/42769

Almeida, G. M. de B., Ferreira, J. P., Correia, M., & Oliveira, G. M. de. (2013). Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. *Estudos Linguísticos*, *42*(1), 204–215.

Almeida, L., Marinho-Araújo, C. M., Amaral, A., & Dias, D. (2012). Democratização do acesso e do sucesso no ensino superior: uma reflexão a partir das realidades de Portugal e do Brasil. *Avaliação*, *17*(3), 899–920. doi:10.1590/S1414-40772012000300014

Alonso, A., Millon, C.,& Williams, G. (2011). Collocational networks and their application to an e-Advanced Learner's Dictionary of Verbs in Science (DicSci). In I. Kosem, & K. Kosem (Eds.), *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex 2011* (pp.12-22). Ljubljana: Trojina, Institute for Applied Slovene Studies.

Antunes, S., Nascimento, M.F.B, Casteleiro, J. M., Mendes, A., Pereira, L., & Sá, T. (2006). A lexical database of Portuguese multiword expressions. In R. Vieira et al. (Orgs.), *PROPOR 2006, LNAI 3960* (pp. 238-243). Berlin: Springer-Verlag.

Atkins, B. T. S., & Varantola, K. (1997). Monitoring dictionary use. *International Journal of Lexicography*, *10*(1), 1–45. doi:10.1093/ijl/10.1.1

Atkins, B. T. S., Clear, J. & Ostler, N. (1991). Corpus design criteria. Retrieved from http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf.

Atkins, B.T.S. (2008). Theoretical lexicography and its relation to dictionary-making. In T. Fontenelle (Ed.), *Practical lexicography. A reader* (pp.31-50). Oxford: Oxford University Press.

Atkins, S., & Grundy, V. (2006). Lexicographic profiling: an aid to consistency in dictionary entry design. In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings XII Euralex International Congress Torino, Italia, September 6 th-9th, 2006* (pp. 1097–1107). Torino: Edizioni dell'Orso.

Baker, P. (2011). Social involvement in corpus studies. In V. Viana; S. Zyngier, & G. Barnbrook (Eds.), *Studies in corpus linguistics: perspectives on corpus linguistics* (pp. 17-27). Amsterdam, NL: John Benjamins Publishing Company.

Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., & Mamede, N. (2010). P-AWL: Academic Word List for Portuguese. In *International Conference on Computational Processing of the Portuguese Language* (pp. 120–123). Berlin Heidelberg: Springer.

Barbara, L. & Macêdo, C.M.M. (2011). Processos verbais em artigos acadêmicos: padrões de realização da mensagem. In L. Barbara & E. Moyano (Orgs.), *Textos e linguagem acadêmica: explorações sistêmicas funcionais em espanhol e português* (pp. 213-231). São Paulo/Campinas: Mercado de Letras.

Barbeiro, L.F., Pereira, L.A., Carvalho, J. B. (2015). Writing at Portuguese universities: students' perceptions and practices. *Journal of Academic Writing*, *5*(1), 74–85.

Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, *16*(1), 3–44. doi:10.1075/ijcl.16.1.02bar

Barros, A. S. X. (2015). Expansão da educação superior no Brasil: limites e possibilidades. *Educação & Sociedade*, *36*(131), 361–390. doi:10.1590/ES0101-7330201596208

Baxter, A. (1992). Portuguese as a pluricentric language. In M. Clyne (Ed.), *Pluricentric languages: differing norms in different nations* (pp. 11-43). Berlin, New York: Mouton de Gruyter.

Bazerman, C., & Moritz, M. E. W. (2016). Higher education writing studies in Latin America. *Ilha do Desterro*, *69*(3), 09–11. Retrieved from

https://periodicos.ufsc.br/index.php/desterro/article/view/2175-8026.2016v69n3p9/32634

Bechara, E. (2001). *Moderna gramática portuguesa*. Rio de Janeiro: Ed. Nova Fronteira.

Benko, V. (2014a). Aranea: yet another family of (comparable) web corpora. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655*. (pp. 257–264). Brno: Springer International Publishing Switzerland. Retrieved from http://www.juls.savba.sk/~vladob

Benko, V. (2014b). Compatible sketch grammars for comparable corpora. In A. Abel & N. Vettori, Chiara Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014.* (pp. 417–430). Bolzano/Bozen: Eurac Research.

Bennett, K. (2008). *English academic discourse: Its hegemonic status and implications for translation* (Doctoral dissertation, University of Lisbon).

Bennett, K. (2010a). Academic discourse in Portugal: a whole different ballgame? *Journal of English for Academic Purposes, 9*(1), 21-32. Retrieved from https://www.academia.edu/1575297/Academic_Discourse_in_Portugal_A_Whole_Different_Ballgame

Bennett, K. (2010b). Academic writing practices in Portugal: survey of Humanities and Social Science researchers. *Diacrítica, 24*(1), 193–209. Retrieved from https://www.academia.edu/1575290/Academic_writing_practices_in_Portugal_survey_of_Humanities_and_Social_Science_researchers

Bergenholtz, H., & Nielsen, J. S. (2013). What is a lexicographical database? *Lexikos, 23*(1), 77-87. Retrieved from http://lexikos.journals.ac.za/pub/article/view/1205

Bezerra, B. G. (2016). A propósito da "síntese brasileira" nos estudos de gêneros. *Revista de Estudos da Linguagem, 24*(2), 465–491. doi:10.17851/2237.2083.24.2.465-491

Bhatia, V. K. (2004). *Worlds of written discourse. A genre-based view*. London and New York: Continuum.

Biber D., Conrad, S., & Leech, G. (2015) *Longman student grammar of spoken and written English*. 14th Impression. Harlow: Pearson.

Biber, D. (1996). Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics, I* (2), 171–197. Retrieved from http://jan.ucc.nau.edu/biber/Biber/Biber_1996.pdf

Biber, D. (2006). *University language. A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamin's Publishing Company.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, *14*(3), 275–311. Retrieved from http://jan.ucc.nau.edu/biber/Biber/Biber_2009.pdf

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*, 263–286. Retrieved from http://jan.ucc.nau.edu/biber/Biber/Biber_Barbieri_2007.pdf

Biber, D., Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. Retrieved from http://jan.ucc.nau.edu/biber/Biber/Biber_Conrad_Cortes_2004.pdf

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.

Bick, E. (2000). *The parsing system Palavras, Automatic grammatical analysis of Portuguese in a constraint grammar framework* (Doctoral dissertation, Aarhus University).

Biderman, M. T. C. (1978). *Teoria lingüística: lingüística quantitativa e computacional*. Rio de Janeiro: Livros Técnicos e Científicos.

Biderman, M. T. C. (1984a). A ciência da lexicografia. *Alfa: Revista de Linguística (São José do Rio Preto)*, *28*(supl.), 1–26.

Biderman, M. T. C. (1984b). O dicionário padrão da língua. *Alfa: Revista de Linguística (São José do Rio Preto)*, *28*(1), 27–43.

Biderman, M. T. C. (1996a). Frequency dictionary of contemporary Portuguese. *Meta*, *41*(2), 275–278.

Biderman, M. T. C. (1996b). Léxico e vocabulário fundamental. *Alfa: Revista de Linguística (São José do Rio Preto)*, *40*, 27–46.

Biderman, M. T. C. (1996c). O dicionário e o vocabulário da língua portuguesa. *Linha D'Água*, *0*(10), 31–39. http://doi.org/10.11606/issn.2236-4242.v0i10p31-39

Biderman, M. T. C. (1998). A face quantitativa da linguagem: um dicionário de freqüências do português. *Alfa: Revista de Linguística (São José do Rio Preto)*, *42*(1), 164–181.

Biderman, M. T. C. (2000). Aurélio: sinônimo de dicionário? *Alfa: Revista de Linguística (São José do Rio Preto)*, *44*, 27–55.

Biderman, M. T. C. (2003a). Análise de dois dicionários gerais do Português Brasileiro contemporâneo: o Aurélio e o Houaiss. *Filologia e linguística portuguesa*, *0*(5), 85–116.

Biderman, M. T. C. (2003b). Dicionários do português: da tradição à contemporaneidade. *Alfa: Revista de Linguística (São José do Rio Preto)*, *47*(1), 53–69.

Biderman, M. T. C. (2005). Unidades complexas do léxico. In G. M. Rio-Torto, O. M. Figueiredo, & F. Silva (Coord.), *Estudos em homenagem ao Professor Doutor Mário Vilela* (pp. 747–758). Porto: Faculdade de Letras da Universidade do Porto. Retrieved from http://ler.letras.up.pt/uploads/ficheiros/4603.pdf

Biderman, M.T. C. (1992). *Dicionário Contemporâneo do Português*. Editora Vozes.

Biewer, C., Nesselhauf, N., & Hundt, M., (2007). *Corpus linguistics and the web*. Amsterdam: Brill Academic Publishers.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *SSLA*, 321–343.

Bogaards, P. (2003). Uses and users of dictionaries. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 26-33). Amsterdam: John Benjamins Publishing.

Bogaards, P. (2010). Lexicography: science without theory? In G.-M. de Schryver (Ed.), *A way with words: recent advances in lexical theory and analysis: a festschrift for Patrick Hanks* (pp. 313-322). Kampala: Menha Publishers.

Bolzan, R. M. (2012). *O uso do dicionário escolar como mediador das práticas discursivas de alunos do ensino fundamental*. (Doctoral dissertation, Universidade Estadual de Londrina).

Borba, F. (Ed.). (2002). *Dicionário de Usos do Português do Brasil.* São Paulo: Ática.

Bortolini, L. S. & Kuhn, T. Z. (2011). Produção de livro didático de nível básico para ensino de português brasileiro para falantes de línguas distantes: decisões teórico-práticas. In R. Teixeira e Silva, Q. Yan, M. A Espadinha, & A. V. Leal (Eds.), *Anais do III SIMELP: A formação de novas gerações de falantes de português no mundo* (pp.13-23). Macau: Universidade de Macau, Departamento de Português.

Branco, A., Mendes, A., Pereira, S., Henriques, P., Pellegrini, T., Meinedo, H., … Bacelar, F. (2012). *The Portuguese Language in the digital era/A língua portuguesa na era digital.* White Paper Series. Springer. Retrieved from http://www.meta-net.eu/whitepapers

Britto, L. P. L., Silva, E. O., Castilho, K. C. de, & Abreu, T. M. (2008). Conhecimento e Formação nas IES Periféricas perfil do aluno "novo" da Educação Superior. *Avaliação*, *13*(3), 777–791.

Calcia, N. P. (2015). O Acordo Ortográfico aplicado aos grafos e aos dicionários do Unitex. *Estudos Linguísticos*, *44*(2), 799–811.

Camargo, M.J.P. d. (2009). *Ensino de português em cursos superiores: razões e concepções*. (Master's thesis, Universidade de Sorocaba).

Camargo, M.J.P. d., & Britto, L.P.L. (2011). Vertentes do ensino de português em cursos superiores. *Avaliação, 16* (2), 345-353.

Cardoso, A. & Magro, C. (2013). Problemas de coesão referencial na escrita académica em português. Paper presented at *3ª Conferência Internacional em Gramática e Texto - GRATO 2013*. Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa, Lisboa.

Cardoso, A., Magro, C., Braz, J. & Nunes, T. (2014). CUTe: Corpus of Portuguese Undergraduates' Texts - Um recurso para a investigação em escrita académica emportuguês. In A. Moreno, F. Silva, I. Falé, I. Pereira & J. Veloso (Orgs.), *XXIX Encontro Nacional da Associação Portuguesa de Linguística - Textos Selecionados* (pp. 169-184). Porto: Associação Portuguesa de Linguística. Retrieved from http://www.cute.org.pt/PT/PUBLICACOES/cardoso_magro_braz_nunes_2014.pdf

Carvalho, J. A. B. (2013). Literacia académica: da escola básica ao ensino superior – uma visão integradora. *Letras & Letras, 29*(2). Retrieved from http://www.seer.ufu.br/index.php/letraseletras/article/view/25983

Casteleiro, J. M. (Coord.). (2001) *Dicionário da língua portuguesa contemporânea*. 2v. Lisboa: Verbo.

Castilho, A. (2005). *Nova gramática do português brasileiro*. São Paulo: Ed. Contexto.

Castilho, A. T., Silva, G. M. O., & Lucchesi, D. (1996). Informatização de acervos da língua portuguesa. In M. F. B. Nascimento, M. C., Rodrigues, & J. B. Gonçalves, (Orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística. (Vol.I) Corpora* (pp. 113–128). Lisboa. Retrieved from http://www.apl.org.pt/docs/actas-11-encontro-apl-1995_vol1.pdf

Cegalla, D. P. (2008). *Novíssima gramática da língua portuguesa.* São Paulo: Ed. Nacional.

Celpe-Bras (Certificado de Proficiência em Língua Portuguesa para Estrangeiro). Retrieved from http://portal.inep.gov.br/celpebras

Census of Higher Education. http://download.inep.gov.br/educacao_superior/censo_superior/apresentacao/2014/coletiva_censo_superior_2013.pdf

Čermák, F. (2003). Source materials for dictionaries. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 18-25). Amsterdam: John Benjamins Publishing.

Certification criteria for SciELO sites: About SciELO (n.d.). Retrieved from http://www.scielo.org/php/level.php?lang=en&component=42&item=3

Charles, M. (2016). All tooled up: Corpus-assisted editing for academic writers. In *Teaching and Language Corpora Conference* (TaLC12) 20-23 July 2016. *Book of Abstracts* (pp. 17–18). Giessen, Germany. Retrieved from http://www.uni-giessen.de/faculties/f05/engl/ling/talc/home/programme/abstracts

Charles, M., Pecorari, D., & Hunston, S. (2009). Introduction: Exploring the interface between corpus linguistics and discourse analysis. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic Writing: At the interface of corpus and discourse* (pp. 1-10). London: Continuum International Publishing Group.

Chen, Q., & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, *26*(4), 502–514. doi:10.1016/j.esp.2007.04.003

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist. 16*(1), 22–29. doi:10.3115/981623.981633

Clear, J. (1987). Overview of the role of computing in Cobuild. In J. Sinclair (Ed.), *Looking Up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary* (pp. 41-61). London: Collins.

Clyne, M. (1992). Pluricentric languages – introduction. In M. Clyne (Ed.), *Pluricentric languages: differing norms in different nations* (pp. 1-9). Berlin, New York: Mouton de Gruyter.

Cock, S. (2002). Pragmatic prefabs in learners' dictionaries. In A. Braasch & C. Povlsen (Eds.), *Proceedings of the 10th EURALEX International Congress* (pp. 471–481). København: Center for Sprogteknologi.

Correia, M. & Ferreira, J. P. (Coords.) (2017). *VOP - Vocabulário Ortográfico do Português*. 2.ª edição. Coimbra: CELGA-ILTEC, Universidade de Coimbra.

Correia, M. & Lemos, L. S. P. (2009). *Inovação lexical em português*. Lisboa: Ed. Colibri.

Correia, M. (1994). Bases digitais lexicais na União Europeia. *Simpósio de lexicologia, lexicografia e terminologia.* Retrieved from http://www.iltec.pt/pdf/wpapers/1994-mcorreia-bdigitais.pdf

Correia, M. (2008). Lexicografia no início do século XXI – novas perspectivas, novos recursos e suas consequências [versão entregue para publicação]. In M.A. Júnior (Coord.), *Lexicon – Dicionário de grego-português, actas de colóquio* (pp. 73-85). Lisboa: Centro de estudos Clássicos / FLUL.

Correia, M. (2009). *Os dicionários portugueses*. Lisboa: Ed. Caminho.

Correia, M., & Guerreiro, P. (1995). Bases de dados lexicais. In M. H. Mateus & A.H. Branco (Orgs.), *Engenharia da linguagem* (pp. 43-69). Lisboa: Ed. Colibri.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397–423. doi:10.1016/j.esp.2003.12.001

Costa, M. C. M.; Rebouças, J. V., & Pontes, A. L. (2014). Representações de professores do curso de letras sobre o dicionário. *Caderno Seminal Digital*, *22* (22), 188-213.

Cowie, A. P. (2001). Homonymy, polysemy and the monolingual English dictionary. *Lexicographica*, *17*, 40–60.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34* (2), 213-238.

Cristovão, V. L. L., & Vieira, I. R. (2016). Literacies in Portuguese and English in Brazilian higher education: Landmarks and perspectives. *Ilha do Desterro*, *69*(3), 209–221. doi:10.5007/2175-8026.2016V69N3P209

Cristóvão, V. L. L., Bork, A. V. B., & Vieira, I. R. (2015). Mapeamento de grupos de pesquisa em torno de letramento (em língua materna): desdobramentos do Projeto ILEES no Brasil. *Letras & Letras*, *31*(3), 73–99. Retrieved from http://www.seer.ufu.br/index.php/letraseletras/article/view/30593

Cruse, D. A. (1986). *Lexical semantics*. New York: Cambridge University Press.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*(1), 66–76. doi:10.1016/j.esp.2013.08.001

de Schryver, G.-M. (2006). Do dictionary users really look up frequent words ? — on the overestimation of the value of corpus-based lexicography. *Lexikos*, *16*, 67–83.

de Schryver, G.-M. (2013). Tools to support the design of a macrostructure. In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with focus on electronic and computational lexicography* (pp. 1384–1395). Berlin: Walter de Gruyter.

de Schryver, G.-M. de, & Prinsloo, D. J. (2000a). Electric corpora as a basis for the compilation of African-language dictionaries. Part 2: the microstructure. *South African Journal of African Languages*, *20*(4), 310–330.

de Schryver, G.-M., & Prinsloo, D. J. (2000b). Electric corpora as a basis for the compilation of African-language dictionaries. Part 1: the macrostructure. *South African Journal of African Languages*, *20*(4), 291–309.

Dias, D. (2015). Has massification of higher education led to more equity? Clues to a reflection on Portuguese education arena. *International Journal of Inclusive Education*, *19*(2), 103–120. doi:10.1080/13603116.2013.788221

Dicionário Eletrônico Houaiss da língua portuguesa (2010). CD-ROM. Editora Objetiva Limitada.

Direção-Geral de Estatísticas da Educação e Ciência (DGEEC). http://www.dgeec.mec.pt/np4/18/

Dobrovoljc, K. (2017). Morphological information in modern Slovene dictionaries. V. Gorjanc, P. Gantar, I. Kosem, & S. Krek (Eds.), *Dictionary of modern Slovene: Problems and solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts.

Dolmacı, M., & Ertaſ, A. (2016). Developing a textbook-based academic Turkish wordlist. *Procedia - Social and Behavioral Sciences*, *232*(April), 821–827. doi:10.1016/j.sbspro.2016.10.111

Duran, M. S. (2008). Métodos na pesquisa de uso de dicionários. *Estudo Linguísticos, 37* (1),31-45.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*(3), 157–169. doi:10.1016/j.esp.2009.02.002

Dutra, D. P., Orfano, B. M., & Sardinha, T. B. (2014). Stance bundles in learner corpora. In S. M. Aluisio & S. E. O. Tagnin (Eds.), *New language technologies and linguistic research: A two-way road* (pp. 2–15). Newcastle upon Tyne: Cambridge Scholars Publishing. Retrieved from http://www.cambridgescholars.com/download/sample/59763

Dziemianko, A., & Lew, R. (2006). Research into dictionary use by Polish learners of English: Some methodological considerations. In K. Dziubalska-Kołaczyk (Ed.), *IFAtuation: A life in IFA. A festschrift for Prof. Jacek Fisiak on the occasion of his 70th birthday by his IFAtuated staff from the School of English, AMU, Poznań* (pp. 211–233). Wydawnictwo UAM. Retrieved from http://www.staff.amu.edu.pl/~rlew/pub/Dziemianko-Lew_Ifatuation_2006.pdf

Eldridge, J. (2008). No, there isn't an "academic vocabulary," but…: A reader responds to K. Hyland and P. Tse's "Is there an "academic vocabulary"?" *TESOL Quarterly*, *42*(1), 109–113. doi:10.1002/j.1545-7249.2008.tb00210.x

Erlandsen, J. (2010). Computational lexicography and lexicology iLEX, a general system for traditional dictionaries on paper and adaptive electronic lexical resources. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress* (p. 306). Leeuwarden/Ljouwert: Fryske Akademy – Afûk.

Evers, A. (2009). *Contraste de verbos indicadores de causalidade em originais e traduções: a Química Geral sob observação*. Retrieved from http://hdl.handle.net/10183/42471.

Ferreira, J. P., Almeida, G. M. de B., & Correia, M. (2013). O uso de corpora para constituição de recursos lexicográficos de referência: o caso do VOC. *Platô, 2*(3), 38–55.

Ferreira, J. P., Barbosa, S., & Janssen, M. (2008). Mordebe Admin: a lexical management system. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress. 15-19 July 2008* (pp. 351–357). Barcelona: Universitat Pompeu Fabra.

Ferreira, J. P., Correia, M., & Almeida, G. de B. (Orgs.) (2017). *Vocabulário Ortográfico Comum da Língua Portuguesa*. Praia: Instituto Internacional da Língua Portuguesa / Comunidade dos Países de Língua Portuguesa.

Ferreira, J. P., Janssen, M., Almeida, G. B. de, Correia, M., & Oliveira, G. M. de. (2012). The Common Orthographic Vocabulary of the Portuguese Language: A set of open lexical resources for a pluricentric language. In *Conference on Language Resources and Evaluation (LREC)* (pp. 1071–1075). Istanbul. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/1034_Paper.pdf

Ferreira, J. P., Lourinho, A., & Correia, M. (2012). Lince, an end user tool for the implementation of the spelling reform of Portuguese. In H. Caseli, A. Villavicencio, A. Teixeira, & F. Perdigão (Eds.), *Computational processing of the Portuguese Language. 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012, Proceedings* (pp. 46–55). Springer.

Finatto, M. J. B., & Huang, C. (2005). Da adjetivação em química e medicina algumas implicações para os estudos do léxico e de textos técnico-científicos. *Revista Língua & Literatura, 6–7*(10–11), 45–56. Retrieved from http://revistas.fw.uri.br/index.php/revistalinguaeliteratura/article/view/38

Finatto, M. J. B., Azerede, S. de, & Lima, E. R. H. de. (2007). Expressões anunciadoras de paráfrase como característica de gêneros textuais: do manual didático de química à legislação ambiental brasileira. In *Simpósio Internacional de Estudos de Gêneros Textuais* (pp. 1472–1482). Tubarão.

Finatto, M. J. B., Evers, A., & Alle, C. M. de O. (2009). Do uso de expressões de causalidade como um elemento caracterizador do gênero textual artigo científico. In *Simpósio Internacional de Estudos de Gêneros Textuais*. Caxias do Sul. Retrieved from http://www.ucs.br/ucs/extensao/agenda/eventos/vsiget/portugues/anais/textos_autor

Finatto, M. J. B., Evers, A., & Alle, C. M. de O. (2010). Para além das terminologias: estudos de convencionalidade em linguagens científicas. In C.L.Perna, H. K. Delgado, & M. J. Finatto (Eds.), *Linguagens especializadas em corpora* (pp. 152–182). Porto Alegre: EDIPUCRS.

Finatto, M. J., Evers, A., Pasqualini, B. F., Kuhn, T. Z & Pereira, A. M. (2014). Vocabulário controlado e redação de definições em dicionários de português para estrangeiros: Ensaios para uma léxico-estatística textual. *Revista Trama*, *10*(20), 53–68.

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal*, *81*(3), 285–300. Retrieved from http://www.jstor.org/stable/329302

Fischer, A. (2008). Letramento acadêmico: uma perspectiva portuguesa. *Acta Scientiarum. Language and Culture*, *30*(2), 177–187. Retrieved from http://periodicos.uem.br/ojs/index.php/ActaSciLangCult/article/view/2334

Fischer, A. (2011). Práticas de letramento acadêmico em um curso de engenharia têxtil: o caso dos relatórios e suas dimensões escondidas. *Scripta*, *15*(28), 37–58.

Fišer, D.& Čibej, J. (2017). The potential of crowdsourcing in modern lexicography. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (Eds.), *Dictionary of modern Slovene: Problems and solutions* (pp.212-228). Ljubljana: Ljubljana University Press, Faculty of Arts.

Flowerdew, J. (2002). Introduction: approaches to the analyses of academic discourse in English. In J. Flowerdew (Ed.), *Academic discourse* (pp. 1-17). Harlow: Longmann.

Flowerdew, L. (2015). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, *20*, 58–68.

Flowerdew, L. (2016). A genre-inspired and lexico-grammatical approach for helping postgraduate students craft research grant proposals. *English for Specific Purposes*, *42*, 1–12. doi:10.1016/j.esp.2015.10.001

Frankenberg-Garcia, A. & Newstad, H. (Eds.). (2015). *Oxford Portuguese Dictionary*. Oxford: Oxford University Press.

Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. *International Journal of Lexicography, 25*(3), 273–296. doi:10.1093/ijl/ecs011

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL, 26*(2), 128-146.doi:10.1017/S0958344014000093

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering automated lexicography: The case of the Slovene lexical database. *International Journal of Lexicography, 29*(2), 200–225. doi:10.1093/ijl/ecw014

Garcia, M., Gamallo, P., Gayo, I., & Cruz, M. A. P. (2014). PoS-tagging the web in Portuguese. National varieties, text typologies and spelling systems. *Procesamiento de Lenguaje Natural, 53*, 95–101.

Gardner, D., & Davies, M. (2016). A response to "To what extent is the Academic Vocabulary List relevant to university student writing?". *English for Specific Purposes, 43*, 62–68. doi:10.1016/j.esp.2016.01.004

Geeraerts, D. (2003). Meaning and definition. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp.83-93). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gorjanc, V., Gantar, P., Kosem, I., & Krek, S. (Eds.). (2017). *Dictionary of modern Slovene: Problems and solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts.

Gouws, R. H. (2003). Types of articles, their structure and different types of lemmata. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp.34-43). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gouws, R. H. (2010). The monolingual specialised dictionary for learners. In P. A. Fuertes-Olivera (Ed.), *Specialised dictionaries for learners* (pp. 55–68). Berlin, Boston : De Gruyter.

Grande Dicionário da Língua Portuguesa (2016). Porto: Porto Editora.

Granger, S. & M. Paquot (2010a). Customising a general EAP dictionary to meet learner needs. In S. Granger & M. Paquot (Eds.), *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009. Cahiers du CENTAL* (pp. 87-96). Louvain-la-Neuve : Presses Universitaires de Louvain.

Granger, S. (2012). Electronic lexicography: From challenge to opportunity. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 1–16). Oxford: Oxford University Press.

Granger, S., & Paquot, M. (2009a). In search of a general academic vocabulary: a corpus-driven study. In K. Katsampoxaki-Hodgetts (Ed.), *Options and practices of LSP practitioners* (pp. 94–108). University of Crete Publications.

Granger, S., & Paquot, M. (2009b). Lexical verbs in academic discourse: A corpus-driven study of expert and learner use. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic Writing: At the Interface of Corpus and Discourse* (pp. 193–214). London: Continuum International Publishing Group.

Granger, S., & Paquot, M. (2010b). The Louvain EAP Dictionary (LEAD). In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress Proceedings of the XIV EURALEX International Congress* (pp.321-326). Leeuwarden/Ljouwert: Fryske Akademy – Afûk

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass, 3*, 1–17. doi:10.1111/j.1749-818X.2009.00149.x

Gries, Stephen. (2011). Methodological and interdisciplinary stance in corpus linguistics. In V. Viana, S. Zyngier, G. Barnbrook (Eds.). *Studies in corpus linguistics: Perspectives on corpus linguistics* (pp. 81-98). Amsterdam, NL: John Benjamins Publishing Company.

Grundy, V. & Rawlinson, D. (2015). The practicalities of dictionary production; planning and managing dictionary projects; training of lexicographers. In P. Durkin (Ed.), *The Oxford handbook of lexicography* (pp. 561-578). Oxford: Oxford University Press.

Haensch, G., Wolf, L., Ettinger, S., & Werner, R. (1982). *La lexicografía. De la lingüística teórica a la lexicografía práctica*. Madrid: Ed. Gredos.

Halliday, M. A. (2004). The language of science. (J. Webster, Ed.), *The collected works of M.A.K. Halliday.* (Volume 5). London: Continuum. doi:10.1017/CBO9781107415324.004

Halliday, M. A., & Martin, J. R. (1996). *Writing science. Literacy and discursive power.* London and Washington, DC: Falmer Press. doi:10.1017/CBO9781107415324.004

Hancioğlu, N., Neufeld, S., & Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes*, *27*(4), 459–479. doi:10.1016/j.esp.2008.08.001

Hanks, P. (1987). Definitions and explanations. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary* (pp. 116-136). London: Collins.

Hanks, P. (2008). The lexicographical legacy of John Sinclair. *International Journal of Lexicography*, *21* (3), 219-229.

Hanks, P. (2009). The impact of corpora on dictionaries. In P. Baker (Ed.), *Contemporary studies in linguistics: contemporary corpus linguistics* (pp. 214–236). London: Continuum.

Hanks, P. (2010). Compiling a monolingual dictionary for native speakers. *Lexikos*, *20*, 580–598. doi:10.4314/lex.v20i1.62738

Hanks, P. (2012a). Corpus evidence and electronic lexicography. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 57–82). Oxford: Oxford University Press. Retrieved from www.patrickhanks.com/uploads/5/1/4/9/5149363/hanks_2012f.pdf

Hanks, P. (2012b). The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), 398–436. doi:10.1093/ijl/ecs026

Hartmann, R. R. K (1989). Sociology of the dictionary user: Hypotheses and empirical studies. In F. J. Hausmann, O. Reichmann, H.E. Wiegand, & L. Zgusta (Eds.), *Wörterbücher, Dictionaries, Dictionnaires. An international encyclopedia of lexicography*, Vol. 1 (pp.102-111). Berlin: Walter de Gruyter.

Hartmann, R. R. K. (1999). What is ‘dictionary research?’. *International Journal of Lexicography*, *12*(2), 155–161. doi:10.1056/NEJMra1313875

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. Harlow: Longman-Pearson Education.

Hartmann, R. R. K. (2008). Twenty-five years of dictionary research: taking stock of conferences and other lexicographic events since Lexeter ’83. In E. Bernal & J. de Cesaris (Eds.), *Proceedings of the XIII Euralex International Congress. 15-19 July 2008* (pp.131-148). Barcelona: Universitat Pompeu Fabra.

Hartmann, R. R. K., & James, G. (1998). *Dictionary of lexicography*. London and New York: Routledge.

Hausmann, F.J., Reichmann, O., Wiegand H. E., & Zgusta, L. (1989). Preface. First Volume. In F. J. Hausmann, O. Reichmann, H.E. Wiegand, & L. Zgusta (Eds.), *Wörterbücher, Dictionaries, Dictionnaires. An international encyclopedia of lexicography* (pp. xvi-xxiv). Berlin, New York: Walter de Gruyter.

Hoey, M. (2005). *Lexical priming: a new theory of words and language*. London and New York: Routledge.

Householder, F. W. (1962). Summary report. In F.W. Householder & S. Saporta (Eds.), *Problems in lexicography* (pp. 279-282). Bloomington: Indiana University.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, *33*(1), 54–65. doi:10.1016/j.esp.2013.07.001

Humblé, P. (2001). *Dictionaries and language learners*. Frankfurt am Main: Haag und Herchen.

Hunston, S. (2006). Corpus linguistics. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (pp. 234–248). Oxford: Elsevier. doi:10.1016/B0-08-044854-2/00944-5

Hyland, K. & Tse, P. (2009). Academic lexis and disciplinary practice: corpus evidence for specificity. *International Journal of English Studies*, *9*(2), p. 111–129. Retrieved from

http://dialnet.unirioja.es/servlet/articulo?codigo=3104270\nhttp://www.eric.Ed.gov/PDFS/EJ878445.pdf

Hyland, K. (2006). *English for Academic purposes. An advanced resource book*. London and New York: Routledge.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62. doi:10.1111/j.1473-4192.2008.00178.x

Hyland, K. (2009). *Academic discourse*. London: Continuum International Publishing Group.

Hyland, K., & Bondi, M. (Eds.). (2006). *Academic discourse across disciplines*. Bern: Peter Lang.

Hyland, K., & Tse, P. (2007). Is there an Academic Vocabulary? *TESOL Quarterly*, *41*(2), 235-253.

Hyon, S. (1996). Genre in Three Traditions: Implications for ESL. *TESOL Quarterly*, 30(4), 693. doi:10.2307/3587930

Jackson, H. (2002). *Lexicography: An introduction*. London and New York: Routledge.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference*. Lancaster (pp. 125–127). Retrieved from http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf

Janssen, M. (2005). Lexical vs. dictionary databases. In F.Kiefer, G. Kiss, & J. Pajzs. (Eds.), *Papers in Computational Lexicography - COMPLEX 2005.* Budapest: Linguistics Institute HAS (pp.1-10).

Jansson, H., Kokkinakis, S. J., Ribeck, J., & Sköldberg, E. (2012). A Swedish academic word list: methods and data. In R.V. Fjeld & J.M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012* (pp. 995–960). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo. Retrieved from http://www.euralex.org/elx_proceedings/Euralex2012/pp955-960 Jansson, Kokkinakis, Ribeck and Skoldberg.pdf

Jezine, E., Chaves, V. L. J., & Cabrito, B. G. (2011). O acesso ao ensino superior no contexto da globalização. Os casos do Brasil e de Portugal. *Revista Lusófona de Educação, 18*, 57–79.

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M.,& Viks, Ü. (2015). Automatic generation of the Estonian collocations dictionary database. In I. Kosem, M. Jakubíček, J. Kallas, & S. Krek (Eds.), *Electronic lexicography in the 21st century: linking lexical data in the digital age.Proceedings of the eLex 2015 conference*. (pp. 1–20). Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. Retrieved from https://elex.link/elex2015/conference-proceedings/

Kelly, J. R. (1970). A computational frequency and range list of five hundred Brazilian-Portuguese words. *Luso-Brazilian Review*, *7*(2), 104–113.

Kiefer, F. & van Sterkenburg, P. (2003). Design and production of monolingual dictionaries. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp.350-365). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kilgarriff, A. (1997). "I don't believe in word senses", *Computers and the Humanities, 31(2),* 91–113.

Kilgarriff, A. (2003). What computers can and cannot do for lexicography or Us precision, them recall. *Keynote lecture at Asialex 2003*. Retrieved from https://www.kilgarriff.co.uk/.../2003-K-AsialexKeynote.doc

Kilgarriff, A. (2006). Collocationality (and how to measure it). In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings XII Euralex International Congress Torino, Italia, September 6 th-9th, 2006* (pp. 997–1004). Torino: Edizioni dell'Orso.

Kilgarriff, A. (2012). Book review of Magali Paquot's Academic vocabulary in learner writing: from extraction to analysis. *International Journal of Corpus Linguistics*, *17*(1), 125–130.

Kilgarriff, A., & Rundell, M. (2002). Lexical profiling software and its lexicographic applications: a case study. In A. Braasch & C. Povlsen (Eds.), *Proceedings of the 10th EURALEX International Congress* (pp. 807–818). København: Center for

Sprogteknologi. Retrieved from http://www.euralex.org/elx_proceedings/Euralex2002/090_2002_V2_Adam Kilgarriff & Michael Rundell_Lexical Profiling Software and its Lexicographic Ap.pdf

Kilgarriff, A., & Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of ACL Workshop on Collocation Computational Xxtraction, Analysis and Exploitation* (pp. 32–38). Retrieved from http://www.kilgarriff.co.uk/Publications/2001-KilgTugwell-ACLcollos-Sketches.pdf

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography: Journal of ASIALEX, 1*, 7–36.

Kilgarriff, A., Baisa, V., Rychlý, P., & Jakubíček, M. (2015). Longest-commonest match. In I. Kosem, M. Jakubíček, J. Kallas, & S. Krek (Eds.). *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference* (pp. 397–404). Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. Retrieved from https://www.sketchengine.co.uk/wp-content/uploads/Longest-commonest_eLex2015.pdf

Kilgarriff, A., Husák, M., Mcadam, K., Rundell, M., & Rychlý, P. (2008). GDEX : automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the 13th EURALEX International Congress* (pp. 425–432). Barcelona: Universitat Pompeu Fabra.

Kilgarriff, A., Jakubícek, M., Pomikalek, J., Sardinha, T. B., & Whitelock, P. (2015). PtTenTen: A corpus for Portuguese lexicography. In T.B. Sardinha & T. L. S. B. Ferreira (Eds.), *Working with Portuguese Corpora* (pp. 111- 127). London, New York: Bloomsbury.

Kilgarriff, A., Kovář, V., & Rychlý, P. (2010). Tickbox lexicography. In S. Granger & M. Paquot (Eds.), *eLexicography in the 21 st Century: New Challenges, New Applications. Proceedings of eLex 2009* (pp. 411–418). Louvain-la-Neuve : Presses Universitaires de Louvain.

Kilgarriff, A., Kovář, V., Krek, S., Srdanovic, I., & Tiberius, C. (2010). A quantitative evaluation of word sketches. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 372–379). Leeuwarden: Fryske Akademy, Afûk.

Kilgarriff, A., Pomikalek, J., Jakubíček, M., & Whitelock, P. (2012). Setting up for corpus lexicography. In R.V. Fjeld & J.M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012* (pp. 778–785). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–115). Lorient : Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Kilian, C. K., & Loguercio, S. D. (2015). Fraseologias de gênero em resumos científicos de linguística, engenharia de materiais e ciências. *TradTerm, 26*, 241–267.

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries. An international encyclopedia of lexicography. Supplement volume: Recent developments with focus on electronic and computational lexicography* (pp. 517–524). Berlin, Boston: De Gruyter.

Kokkinakis, S. J., Sköldberg, E., Henriksen, B., Kinn, K., & Johannessen, B. J. (2012). Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project. In R.V. Fjeld, & J.M. Torjusen, (Eds.), *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012* (pp.563-569). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo

Koplenig, A. & Müller-Spitzer, C. (2014). Questions of design. In C. Müller-Spitzer (Ed.), *Using online dictionaries (Lexicographica: Series maior - 145)* (pp. 189–204). Berlin, Boston: De Gruyter.

Koplenig, A. (2014). Empirical research into dictionary use. In C. Müller-Spitzer (Ed.), *Using online dictionaries (Lexicographica : Series maior - 145)* (pp. 55-76). Berlin, Boston: De Gruyter.

Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks. [Automatic detection of good dictionary examples in Estonian learner's dictionaries.]. *Eesti Rakenduslingvistika Ühingu Aastaraamat [Estonian Papers in Applied Linguistics]*, *13*, 53–71.

Kosem, I.; Koppel, K.; Kuhn, T. Z.; Michelfeit, J.; Tiberius, C. (forthcoming). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*.

Kosem, I., Gantar, P., & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Eds.), *Electronic Lexicography in the 21st Century: Thinking Outside the Paper: Proceedings of the eLex 2013 Conference* (pp. 32–48). Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=4561372&info=resumen&idioma=ENG

Kosem, I., Gantar, P., Logar, N. & Krek, S. (2014). Automation of lexicographic work using general and specialized corpora: two case studies. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus* (pp. 355–364). Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.

Kosem, I., Husák, M., & McCarthy, D. (2011). GDEX for Slovene. In I. Kosem, & K. Kosem (Eds.), *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex 2011* (pp. 151–159). Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=4567470&info=resumen&idioma=ENG

Kosem, Iztok. (2010). *Designing a model for a corpus-driven dictionary of academic English*. (Doctoral dissertation, Aston University).

Koskela, A. (2015). Identification of homonyms in different types of dictionaries. In P. Durkin (Ed.), *The Oxford handbook of lexicography* (pp.457-471). Oxford: Oxford University Press.

Krieger, M. da G. (2015). O léxico com letra maiúscula: reflexo do trabalho de Maria Tereza Biderman. *Debate Terminológico*, *14*, 89–91.

Krishnamurthy, R. (1987). The process of compilation. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary* (pp. 62-85). London: Collins.

Kuhn, T. Z., & Kosem, I. (2016). Devising a sketch grammar for academic Portuguese. *Slovenščina 2.0, 4* (1), 124–161.

Kuhn, T. Z. (2015). Resenha de Oxford Learner's Dictionary of Academic English. *BELT+Brazilian English Language Teaching Journal*, *6*(1), 110–119.

Kuhn, T. Z., & Ferreira, J. P. (2016). Building a corpus of written academic texts in Portuguese. In *Teaching and Language Corpora Conference (TaLC12) 20-23 July 2016. Book of Abstracts* (p.103). Giessen, Germany. Retrieved from http://www.uni-giessen.de/faculties/f05/engl/ling/talc/home/programme/abstracts

Kuhn, T. Z., & Finatto, M. J. B. (2011). On the proposal of an on-line Brazilian Portuguese dictionary for speakers of Asian languages: an ongoing experiment. In K. Akasu & S. Uchida (Eds.), *Lexicography: Theoretical and practical perspectives. ASIALEX2011 Proceedings* (pp. 284-293). Kyoto: The Asian Association for Lexicography.

Kuhn, T. Z., Finatto, M. J. B., & Evers, A. (2011). Uso de vocabulário controlado em dicionários de português como língua estrangeira em formato on-line: uma experiência em andamento para uso de aprendizes coreanos. In R. Teixeira e Silva, Q. Yan, M. A Espadinha, & A. V. Leal (Eds.), *Anais do III SIMELP: A formação de novas gerações de falantes de português no mundo* (pp. 45-57). Macau: Universidade de Macau, Departamento de Português.

Landau, S. (2001). *The art and craft of lexicography*. Cambridge: Cambridge University Press.

Lea, D. (2014). Making a learner's dictionary of academic English. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX international congress: the user in focus* (pp. 181-190). Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*, 56–75. doi:10.1016/j.esp.2005.02.010

Lee, D.Y. W. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5*(3), 37–72.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics*. London: Longman (pp.8-29). Retrieved from http://ccl.pku.edu.cn/doubtfire/CorpusLinguistics/Introduction/The%20state%20of%20the%20art%20in%20corpus%20linguistics.htm

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics: proceedings of Nobel symposium 82* (pp. 105–122). Berlin, New York: Mouton de Gruyter.

Leech, G. (2001). The role of frequency in ELT: new corpus evidence brings a re-appraisal [manuscript]. In W. Hu (Ed.), *ELT in China 2001: Papers presented at the 3rd International Symposium on ELT in China* (pp. 1-23). Beijing: Foreign Language Teaching and Research Press. Retrieved from http://www.lancaster.ac.uk/fass/doc_library/linguistics/leechg/leech_2001.pdf

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22*, 42–53. doi:10.1016/j.jeap.2016.01.008

Lew, R. (2004). *Which dictionary for whom? Receptive use of monolingual, bilingual and semi-bilingual dictionaries by Polish learners of English*. Poznan: Motivex.

Lew, R. (2010a). Multimodal lexicography: the representation of meaning in electronic dictionaries. *Lexicographica, 20*, 290–306.

Lew, R. (2010b). New ways of indicating meaning in electronic dictionaries: hope or hype? In *Learner's lexicography and second language teaching* (manuscript, pp. 1–12). Retrieved from http://www.staff.amu.edu.pl/~rlew/pub/Lew_New_ways_of_indicating_meaning.pdf

Lew, R. (2011). Space restrictions in paper and electronic dictionaries and their implications for the design of production dictionaries. In P. Bański & B. Wójtowicz (Eds.), *Issues in modern lexicography*. München: Lincom Europa. Retrieved from http://works.bepress.com/robert_lew/1/

Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (Ed.), *The Bloomsbury companion to lexicography* (Preprint v, pp. 1–15). London: Bloomsbury Publishing. Retrieved from http://www.bloomsbury.com/uk/bloomsbury-companion-to-lexicography-9781441145970/

Lew, R. (2015a). Dictionary and their users. In P. Hanks, & G-M. de Schryver (Eds.), *International handbook of modern lexis and lexicography* (pp.1-9). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://www.staff.amu.edu.pl/~rlew/pub/Lew_2015_Dictionaries_and_Their_Users.pdf

Lew, R. (2015b). Opportunities and limitations of user studies. *OPAL - Online Publizierte Arbeiten Zur Linguistik, (2)*, 6-16. Retrieved from http://pub.ids-mannheim.de/laufend/opal/pdf/opal15-2.pdf#page=6

Lew, R., & de Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, *27*(4), 341–359.

Lew, R., & Doroszewska, J. (2009). Electronic dictionary entries with animated pictures: Lookup preferences and word retention. *International Journal of Lexicography*, *22*(3), 239–257. doi:10.1093/ijl/ecp022

Lipka, L. (1990). *An outline of English lexicology: lexical structure, word semantics, and word-formation*. Tübingen: De Gruyter.

Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, *39*, 1–11. doi:10.1016/j.esp.2015.03.001

Logar, N. (2014). A corpus based e-dictionary of terminology as a body of knowledge. In Budin, G. (Ed.), *Languages for special purposes in a multilingual, transcultural world proceedings of the 19th European Symposium on Languages for Special*

Purposes, LSP, 2013 8-10 July 2013 Vienna, Austria (pp. 386–392). Vienna: Centre for Translation Studies.

Logar, N., & Kosem, I. (2013). TERMIS: a corpus-driven approach to compiling an e-dictionary of terminology. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Eds.), *Electronic Lexicography in the 21st Century: Thinking Outside the Paper: Proceedings of the eLex 2013 Conference* (pp. 164–178). Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Logar, N., Gantar, P., & Kosem, I. (2014). Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0*, *1*, 41–61.

Longman Dictionary of Contemporary English, 1st edition (LDOCE1) (1978). Harlow: Longman.

Lousada, E. G., Ferreira, A. D., Bueno, L., Rojo, R., Aranha, S., & Abreu-Tardelli, L. (Eds.). (2016). *Diálogos brasileiros no estudo de gêneros textuais/ discursivos*. Araraquara: Letraria.

Lugli, V. C. P. & Silva, M. M. A. (2012). Dicionário: um material didático por excelência no ensino de língua portuguesa para aprendizes surdos. *Anais do X encontro do CELSUL – Círculo de Estudos Linguísticos do Sul UNIOESTE* - Universidade Estadual do Oeste do Paraná.

Lyons, J. (1981). *Linguagem e lingüística*. São Paulo: LTC.

Macmillan English Dictionary for Advanced Learners. Retrieved from http://www.macmillandictionary.com/

Marinho, M. (2010). A escrita nas práticas de letramento acadêmico. *Revista Brasileira de Linguística Aplicada*, *10* (2), 363–386.

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3), 183–198. doi:10.1016/j.esp.2009.04.003

Mateus, M.H. & Branco, A.H. (Eds.). (1995). *Engenharia da linguagem*. Lisboa: Ed. Colibri.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics: an introduction*. Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Abingdon: Routledge.

Meyer, C. F. (2002). *English corpus linguistics: an introduction*. Cambridge: Cambridge University.

Ming-Tzu, K.W., & Nation, P. (2004). Word meaning in academic English: Homography in the academic word list. *Applied Linguistics*, *25*(3), 291-314.

Moerdijk, F. (2003). The codification of semantic information. In P. van Sterkenburg, (Ed.), *A practical guide do lexicography* (pp. 273-296). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Molsing, K. V., & Perna, C. B. L (2014). Research and teaching in Portuguese for Specific Purposes. *BELT +Brazilian English Language Teaching Journal*, *5*(2), 1-7. Retrieved from http://revistaseletronicas.pucrs.br/ojs/index.php/belt/article/view/19701/12527

Montanari, F.& Packer, A.L. (2014). Criteria for the selection of journals to index and publish in the SciELO network collections. In A. L. Packer, N. Cop, A. Luccisano, A. Ramalho, & E. Spinak (Eds.), *SciELO: 15 years of open access - an analytic study of open access and scholarly communication* (pp.67-80). UNESCO. Retrieved from doi:10.7476/9789230012373.

Moon, R. (1998). On using spoken data in corpus lexicography. In T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin & S. Theissen (Eds.), *Proceedings of the 8th EURALEX International Congress* (pp. 347-355). Liège: EURALEX.

Moon, R. (2000). Lexicography and disambiguation: the size of the problem. *Computers and the Humanities*, *34*, 99–102.

Moon, R. (2007). Sinclair, lexicography, and the Cobuild Project. *International Journal of Corpus Linguistics*, *2*(12), 159–181.

Moon, R. (2008). Sinclair, phraseology, and lexicography. *International Journal of Lexicography*, *21*(3), 243–254.

Morais, F. B. C. (2014). Os dizentes nos artigos científicos de linguística – um estudo baseado na Linguística Sistêmico-Funcional e com o auxílio da Linguística de Corpus. *Letras & Letras*, *30*(2), 46–63. Retrieved from http://www.seer.ufu.br/index.php/letraseletras/article/view/27810

Morais, F. B. C. (2015). O uso do processo existencial haver na escrita acadêmica – um estudo com base em um corpus de artigos científicos de diversas áreas do conhecimento. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *27*(1), 142–160. Retrieved from http://www.periodicos.ufes.br/contextoslinguisticos/article/view/10768

Morais, F. B. C. (2016a). Ergatividade x transitividade: um estudo em construções médias em artigos científicos de diferentes áreas do conhecimento. *Domínios de Lingu@gem*, *10*(1), 183–201.

Morais, F. B. C. (2016b). Variação de usos do clítico na comunidade acadêmica: um estudo descritivo com base na linguística sistêmico-funcional. *Cadernos de Linguagem e Sociedade*, *17*(1), 70–100. Retrieved from http://www.periodicos.unb.br/index.php/les/article/view/16078

Morais, F.B.C. (2013). *Entre alhos e bugalhos – os usos de clítico se na escrita acadêmica*. (Doctoral dissertation, PUCSP).

Moro, B. I. (2014). *Advérbios de posicionamento em textos escritos de português acadêmico*. (Master's thesis, PUCRS).

Motta-Roth, D. (1999). A importância do conceito de gêneros discursivos no ensino de redação acadêmica. *Intercâmbio. Revista do Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem*, *8*, 119–129. Retrieved from http://revistas.pucsp.br/index.php/intercambio/article/view/4029

Motta-Roth, D. (2008). Análise crítica de gêneros: contribuições para o ensino e a pesquisa de linguagem. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *24*(2), 341–383. doi:10.1590/S0102-44502008000200007

Motta-Roth, D. (2009). The role of context in academic text production and writing pedagogy. In C. Bazerman, A. Bonini, & D. Figueiredo (Eds.), *Genre in a changing world: Perspectives on writing* (pp. 317–336). Fort Collins, Colorado: The WAC

Clearinghouse and Parlor Press. Retrieved from http://wac.colostate.edu/books/genre/

Motta-Roth, D. (2012). Academic literacies in the South: Writing practices in a Brazilian university. In C. Thaiss, G. Bräuer, P. Carlino, L. Ganobcsik-Williams, & A. Sinha (Eds.), *Writing programs worldwide: Profiles of academic writing in many places* (pp. 105–116). Colorado State University, EUA: The WAC Clearinghouse.

Motta-Roth, D., & Barbara, L. (2012). Foreword. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 28*(SPE), vii–ix. doi:10.1590/S0102-44502012000300001

Motta-Roth, D., & Heberle, V. M. (2015). A short cartography of genre studies in Brazil. *Journal of English for Academic Purposes*, *19*, 22–31.

Motta-Roth, D., & Hendges, G. R. (2010). *Produção textual na universidade*. São Paulo: Parábola.

Muller, J. (2011). Through others' eyes: The fate of disciplines. In F. Christie, & K. Maton (Eds.), *Disciplinarity: functional linguistic and sociological perspectives* (pp. 13-34). London: Continuum.

Müller-Spitzer, C. (2013). Contexts of dictionary use. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Eds.), *Electronic Lexicography in the 21st Century: Thinking Outside the Paper: Proceedings of the eLex 2013 Conference* (pp. 6-13). Liublliana [Eslovenia]: Institute for Applied Slovene Studies; Tallin [Estonia]: Institute of the Estonian.

Müller-Spitzer, C. (Ed.). (2014). *Using online dictionaries (Lexicographica Series Maior 145)*. Berlin: De Gruyter.

Murakawa, C. de A. A. (2001). Resenha Dicionário Didático do Português de Maria Tereza Camargo Biderman. *Todas as Letras*, *3*, 93–97.

Nascimento, M. F. B. (1998). Dicionário de combinatórias do português. *Alfa: Revista de Linguística (São José do Rio Preto)*, *42*(n.esp), 183–203.

Nascimento, M. F. B., Marques M. L.G, & Cruz, M. L. S. (1987). *Português fundamental, métodos e documentos*, tomo 1, Inquérito de Frequência, Lisboa, INIC, CLUL.

Nascimento, M. F. B., Rivenc, P.M.L. & Cruz, M. L. S. (1987). *Português fundamental, métodos e documentos*, tomo 2, Inquérito de Disponibilidade, Lisboa, INIC, CLUL.

Nascimento, M.F. & Gonçalves, J. B. (1996). Corpus de Referência do Português Contemporâneo (CRPC) - desenvolvimento e aplicações. In M. F. B. Nascimento, M. C., Rodrigues, & J. B. Gonçalves (Orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística. (Vol.I) Corpora* (pp. 143-150). Lisboa. Retrieved from http://www.apl.org.pt/docs/actas-11-encontro-apl-1995_vol1.pdf

Nascimento, M.F.B., Rodrigues, M. C., & Gonçalves, J. B. (1996). Corpora portugueses. In M. F. B. Nascimento, M. C., Rodrigues, & J. B. Gonçalves (Orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística. (Vol.I) Corpora* (pp. 423–447). Lisboa. Retrieved from http://www.apl.org.pt/docs/actas-11-encontro-apl-1995_vol1.pdf

Nation, P, & Kyongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*(1), 35-41.

Nation, Paul. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesi, H. (2012). Dictionary use. *The encyclopedia of applied linguistics*, *(November 2012)*, 1–5. doi:10.1002/9781405198431.wbeal0317

Nesi, H. (2014). Dictionary use by English language learners. *Language Teaching*, *47*(01), 38–55. doi:10.1017/S0261444813000402

Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at British university. *International Journal of Lexicography*, *15*(4), 277–305.

Neves, M. H. M. (2011). Gramática: reflexões sobre um percurso de elaboração de manuais introdução. *Revista da ABRALIN*, *(n.Especial)*, 33–51.

Neves, M. H. M. (2010). O acordo ortográfico da língua portuguesa e a meta de simplificação e unificação. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *26*(1), 87–113.

Ninin, M. O. G. (2014). Pode ser... Poderia ser... O uso de modalizações na escrita acadêmica. *Revista Brasileira de Linguística Aplicada*, *14*(1), 175–197. Retrieved from /scielo.php?script=sci_arttext&pid=&lang=pt

Ninin, M. O. G. (2015). Escrita acadêmica e gramática sistêmico-funcional: perspectivas para o ensino. *Trabalhos em Linguística Aplicada*, *54*(3), 593–619. doi:10.1590/010318134658164921

Ninin, M. O. G., & Barbara, L. (2013). Engajamento na perspectiva línguistíca sistêmico-funcional em trabalhos de conclusão de curso de letras. *Trabalhos em Linguística Aplicada*, *52*(1), 127–146.

Ninin, M. O. G., Joseph, N. L. de L., & Maciel, A. M. C. (2015). Metáforas gramaticais como recurso para empacotamento no texto acadêmico. *Letras*, *0*(50), 207–230.

Nunes, E. (2007). Desafio estratégico da política pública: o ensino superior brasileiro. *Revista de Administração Pública*, *41*(edição especial), 1967–2007.

Nunes, S. & Perna, C. B. L. (2015). Uma análise dos substantivos como marcadores de posicionamento em artigos acadêmicos em língua portuguesa. *BELT +Brazilian English Language Teaching Journal*, *6* (2) (julho/dezembro), 172-195. doi:10.15448/2178-3640.2015.2.22635

Oakey, D. (2005). Academic vocabulary in academic discourse: the phraseological behaviour of EVALUATION in economics research articles. In E. Tognini-Bonelli, G. L. Camiciotti (Eds.), *Strategies in academic discourse*. Amsterdam: John Benjamins Publishing.

ODAE. *Oxford Dictionay of Academic English* (2015). Oxford: Oxford University Press

Oliveira, G. M. de. (2013). Política linguística e internacionalização: a língua portuguesa no mundo globalizado do século XXI. *Trabalhos em Linguística Aplicada*, *55* (2) (jul. /dez.), 409–443.

Packer, A. L., Cop, N., Luccisano, A., Ramalho, A., & Spinak, E. (Eds.) (2014). *SciELO: 15 years of open access- an analytic study of open access and scholarly communication*. UNESCO. Retrieved from doi:10.7476/9789230012373.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA* (pp. 1–7). Istanbul. Retrieved from http://nlp.lsi.upc.edu/publications/papers/padro12.pdf

Paquot, M. (2007). Towards a productively-oriented word list. In J. Walinski, K. Kredens, S. Gozdz-Roszkowski (Eds), *Corpora and ICT in Language Studies. PALC 2005* (pp. 127-140). Lodz Studies in Language 13. Frankfurt am Main: Peter Lang. Retrieved from: http://sites-test.uclouvain.be/cecl/archives/PAQUOT_2007_Towards_a%20productively-oriented_academic_word_list.pdf

Paquot, M. (2010). *Academic vocabulary in learner writing. From extraction to analysis*. London, New York: Continuum International Publishing Group.

Paquot, M. (2012). The LEAD dictionary-cum-writing aid: An integrated dictionary and corpus tool. In S. Granger, & M. Paquot (Eds.), *Eletronic lexicography* (pp. 161-186). Oxford: Oxford University Press.

Peixoto, R. M. T. (2015). *O fenômeno (de)queísta no corpus do português brasileiro acadêmico*. (Master's thesis, PUCRS).

Pereira, L. A., & Graça, L. (2014). Concepções sobre a escrita académica de estudantes do ensino superior. *Raído, 8*(16), 125–139. Retrieved from http://www.periodicos.ufgd.edu.br/index.php/Raido/article/viewFile/3751/2019

Perini, M. A. (2002). *Modern Portuguese: A reference grammar*. New Haven: Yale University Press.

Plano de Ação de Lisboa (PALis). Retrieved from http://www.elo-online.org/portal/images/stories/docs/resolucaolisboa.pdf

Ramos, L. B., & Finatto, M. J. B. (2012). Padrões verbais em textos de pediatria: contrapontos com o registro dicionarizado. Retrieved from http://www.nilc.icmc.usp.br/elc-ebralc2012/anais/andamento/103461.pdf

Raposo, E. B. P., Nascimento, M.F.B, Mota, M.A.C., Segura, L., & Mendes, A. (Orgs.) (2013) *Gramática do português*, vol. I and II. Lisboa: Fundação Calouste Gulbenkian.

Rinck, F., Silva, J. Q. G., & Assis, J. A. (2012). Qual abordagem erigir para pensar as práticas de leitura e escrita na formação acadêmica e/ou na vida profissional? *Scripta*, *16*(30), 7–15. Retrieved from http://periodicos.pucminas.br/index.php/scripta/article/view/4255

Rivenc, P. (1996). Réalisme et utopie. Quelques réflexions d'un vieux routard. In M. F. B. Nascimento, M. C., Rodrigues, & J. B. Gonçalves, (Orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística. (Vol.I) Corpora* (pp. 21–38). Lisboa. Retrieved from http://www.apl.org.pt/docs/actas-11-encontro-apl-1995_vol1.pdf

Rodrigues, L. C. S. (2009). *Dificuldades de síntese na escrita de alunos do ensino superior politécnico*. (Doctoral dissertation, Universidade de Aveiro). Retrieved from http://hdl.handle.net/10773/1484

Rundell, M. (1998). Recent trends in English pedagogical lexicography. *International Journal of Lexicography*, *11*(4), 315-342.

Rundell, M. (1999). Dictionary use in production. *International Journal of Lexicography*, *12*(1), 35–53. doi:10.1093/ijl/12.1.35

Rundell, M. (2009). The road to automated lexicography: First banish the drudgery... then the drudges? In S. Granger, & M. Paquot (Eds.), *eLexicography in the 21st century: New challenges, new applications eLex 2009*. *Book of Abstracts* (pp. 9–10).

Rundell, M. (Ed.). (2010). *Macmillan Collocations Dictionary*. Oxford: Macmillan Education.

Rundell, M., & Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora : In honour of Sylviane Granger* (pp. 257-281). Amsterdam: John Benjamins.

Salas Valdebenito, M. (2015). Una propuesta de taxonomía de marcadores metadiscursivos para el discurso académico-científico escrito en español. *Revista Signos*, *48*(87), 95-120. doi:10.4067/S0718-09342015000100005

Sampson, J. (1997). "Genre", "style" and "register". Sources of confusion ? *Revue Belge de Philologie et D'histoire*, *75*(3), 699–708.

Santos, D. (1999). *Comparação de corpora em português: algumas experiências*. Retrieved from http://www.linguateca.pt/documentos/CCP.pdf

Santos, D. (2011). Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *Language Variation Infrastructure*, *3*(2), 113–128.

Santos, D. (2015). Corpora at Linguateca: Vision and roads taken. In T.B. Sardinha & T. L. S. B. Ferreira (Eds.), *Working with Portuguese Corpora* (pp. 219-236). London, New York: Bloomsbury.

Santos, J. V., & Silva, P. N. (2016). Issues of textual hybridity in a major academic genre: PhD dissertations vs research articles. *REDIS: Revista de Estudos Do Discurso*, *5*(January), 171–194.

Santos, S. N. (2015). *Uma Análise dos Substantivos como Marcadores de Posicionamento em Artigos Acadêmicos em Língua Portuguesa*. (Doctoral dissertation, PUCRS).

Santos, S.M. & Packer, A.L. (2014). Production of SciELO Collections and Journals. In A. L. Packer, N. Cop, A. Luccisano, A. Ramalho, & E. Spinak (Eds.), *SciELO: 15 years of open access - an analytic study of open access and scholarly communication* (pp.81-92). UNESCO. Retrieved from doi:10.7476/9789230012373.

Sardinha, T. B. (2000). Lingüística de corpus: histórico e problemática. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *16*, 323–367.

Sardinha, T. B. (2004). *Lingüística de corpus*. Barueri: Manole.

Sardinha, T. B. (2007). Metaphor in corpora: a corpus-driven analysis of applied linguistics dissertations. *Revista Brasileira de Linguística Aplicada*, *7*(1), 11–35. Retrieved from http://www.scielo.br/pdf/rbla/v7n1/02.pdf

Sardinha, T.B., Teixeira, R. B. S., & Ferreira, T:L.S.B (2015). Lexical bundles in Brazilian Portuguese. In T.B. Sardinha & T. L. S. B. Ferreira (Eds.), *Working with Portuguese Corpora* (pp. 33-67). London, New York: Bloomsbury.

Scarton, C.E., & Aluísio, S.M. (2010). Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, *2*(1), 45-61.

Schlatter, M., & Garcez, P. (2009). Línguas Adicionais (Espanhol e Inglês). In *Referenciais curriculares do Estado do Rio Grande do Sul: linguagens, códigos e suas tecnologias* (pp.127-172). Rio Grande do Sul, Secretaria do Estado da Educação, Departamento Pedagógico. Porto Alegre: SE/DP.

Schlatter, M.; Garcez, P. M., & Scaramucci, M. V. R. (2004). O papel da interação na pesquisa sobre aquisição e uso de língua estrangeira: Implicações para o ensino e para a avaliação. *Letras de Hoje*, *39*(3), 345–378.

Seretan, V., & Wehrli, E. (2013). Context-sensitive look-up in electronic dictionaries. In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography* (pp. 1046–1052). Berlin, New York: De Gruyter.

Shepherd, T. M. G. (2014). Changing 'faces': A case study of complex prepositions in Brazilian Portuguese. In T.B. Sardinha & T. L. S. B. Ferreira (Eds.), *Working with Portuguese corpora* (pp. 69-85). London, New York: Bloomsbury.

Silva, P. N. & Santos, J. V. (2011). Contributos para a caracterização do género académico "resposta de desenvolvimento." In R. Teixeira e Silva, Q. Yan, M. A Espadinha, A. V. Leal (Eds.), *Anais do III SIMELP: A formação de novas gerações de falantes de português no mundo* (pp. 27–38). Macau: Universidade de Macau, Departamento de Português.

Silva, P. N. & Santos, J. V. (2015). Da Introdução ao Resumo / Abstract: o surgimento de um género híbrido nas atas da Associação Portuguesa de Linguística. In *Estudos Linguísticos, 10*, 313–336.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics* (Prepublication draft). Retrieved from https://www.google.pt/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwjArMni44TWAhUHVxoKHdOWBl4QFgglMAA&url=http%3A%2F%2Fwww-personal.umich.edu%2F~ncellis%2FNickEllis%2FPublications_files%2FAFL_paper_AppLinxPrepub.pdf&usg=AFQjCNGW3EeRWtro_fSbXvpgEkQi1TN7oQ

Sinclair, J. & Hanks, P. (1987). *Collins COBUILD Advanced Learner's English Dictionary (COBUILD).* London, Glasgow: Collins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1996). Tipologia textual EAGLES. In M. F. B. Nascimento, M. C., Rodrigues, & J. B. Gonçalves, (Orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística. (Vol.I) Corpora* (pp. 39–91). Lisboa. Retrieved from http://www.apl.org.pt/docs/actas-11-encontro-apl-1995_vol1.pdf

Sinclair, J. (2003a). Corpora for lexicography. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp.167-178). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, J. (2003b). Corpus processing. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp.179-193). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp.1-16). Oxford: Oxbow Books. Retrieved from http://ahds.ac.uk/linguistic-corpora/

Sinclair, J. (Ed.) (1987). *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. London: Collins.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149–183. doi:10.1177/0741088316631527

Stubbs, M. (2006). Corpus analysis: the state of the art and three types of unanswered questions. In G. Thompson & S. Hunston (Eds.), *Functional linguistics: System and corpus: Exploring connections* (pp. 15–36). Bristol: Equinox Publishing Ltd.

Summers, D. (1993). Longman / Lancaster English Language Corpus - criteria and design. *International Journal of Lexicography*, *6*(3), 181-208.

Svénsen, B. (2009). *A handbook of lexicography. The theory and practice of dictionary-making.* Cambridge: Cambridge University Press.

Swales, J. (1990). *Genre analysis: English in academic and researching settings*. Cambridge: Cambridge University Press.

Swales, J. (2004). *Research genres. Explorations and applications*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524827

Swales, J. M. (1997). English as Tyrannosaurus rex. *World Englishes*, *16*(3), 373–382. doi:10.1111/1467-971X.00071

The Modern Language Journal. (1997). *Focus issue: Second Language Acquisition Reconceptualized? The Impact of Firth and Wagner* (1997). Retrieved from https://www.jstor.org/stable/i413248

Thomas, J. (2014). *Discovering English with Sketch Engine*. Brno:Versatile

Tiberius, C., Heylen, K., & Krek, S. (2015). *Automatic acquisition of knowledge survey results.* Retrieved from http://www.elexicography.eu/working-groups/working-group-3/wg3-meetings/wg3-herstmonceux-2015/

Todd, R. W. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, *45*, 31–39. doi:10.1016/j.esp.2016.08.003

Tognini-Bonelli, E. (2001). The corpus-driven approach. *Studies in corpus linguistics: Corpus linguistics at work* (pp. 84-100). Amsterdam: John Benjamins.

Torres, L. S., Rodrigues, R., & Aluísio, S. M. (2014). Espanhol-Acadêmico-Br: a corpus of academic Portuguese learners produced by native speakers of Spanish. In S. M. Aluisio & S. E. O. Tagnin (Eds), *New language technologies and linguistic research: A two-way road* (pp. 98-111). Newcastle upon Tyne: Cambridge Scholars.

Trap-Jensen, L. (2010). One, two, many: customization and user profiles in internet dictionaries. In A. Dykstra & A. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 1133–1143). Leeuwarden/Ljouwert: Fryske Akademy – Afûk.

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes, 12*(4), 248–263. doi:10.1016/j.jeap.2013.07.001

van Sterkenburg, P. (2003). 'The' dictionary: Definition and history. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 3-17). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Varantola, K. (2002). Use and usability of dictionaries: common sense and context sensibility? In M.-H. Corréard (Ed.), *Lexicography and natural language processing: A festschrift in honour of BTS Atkins* (pp. 30–44). Stuttgart: Euralex.

Varantola, K. (2003). Linguistic corpora (databases) and the compilation of dictionaries. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 228-239). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ventura, A. S. (2014). *O suporte digital como fonte de informação: um estudo de caso sobre o uso do dicionário, com alunos participantes do projeto UCA* (BA thesis, Universidade Federal do Rio Grande do Sul).

Viana, V., Zyngier, S., & Barnbrook, G. (Eds) (2011). *Perspectives on corpus linguistics.* Amsterdam/Philadelphia: John Benjamins.

Vilela, M. (1990). *Dicionário do Português Básico*. Porto: Asa, 1990

Welker, H. A. (2004). *Dicionários. Uma pequena introdução à lexicografia.* (2.Ed.). Brasília: Thesaurus Editora.

Welker, H. A. (2006). Pesquisando o uso de dicionários. *Linguagem & Ensino, 9*(2), 223–243.

Welker, H. A. (2010). *Dictionary use: A general survey of empirical studies*. Brasília: Author's Edition.

Werner, R. (1982). La definición lexicográfica. In G. Haensch, L. Wolf, S. Ettinger, & R. Werner. *La lexicografía. De la lingüística teórica a la lexicografía práctica* (pp. 259-328). Madrid: Editorial Gredos.

Wynne, M. (Ed). (2005). *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books. Retrieved from http://ahds.ac.uk/linguistic-corpora/

Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes*, *37*, 27–38. doi:10.1016/j.esp.2014.05.003

Zgusta, L. (1971). *Manual of lexicography* (Janua Linguarum Series Maior 39). Prague: Academia. / The Hague: Mouton.

Zgusta, L., Dolezal, F. F. M., & Creamer, T. B. I (2006). *Ladislav Zgusta. Lexicography then and Now. Selected essays*. (*Lexicographica. Series Maior*). Berlin/Boston: De Gruyter.

Generation Lusophonia: why is Portuguese the new language of power and trade. (2012). *Monocle, 56*. Retrieved from http://monocle.com/magazine/issues/57/

Brasil. Ministério da Educação. Institui o Programa Idiomas sem Fronteiras e dá outras providências. Portaria nº 973, de 14 de novembro de 2014. Diário Oficial da União, Nº 222, 17 de novembro de 2014, Seção 1, página 11. Retrieved from http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=17/11/2014&jornal=1&pagina=11

Ferreira, J. P.; Correia, M.; Almeida, G. B. (Orgs.) *Vocabulário Ortográfico Comum da Língua Portuguesa*. Praia: Instituto Internacional da Língua Portuguesa / Comunidade dos Países de Língua Portuguesa. Retrieved from

Acordo Ortográfico da Língua Portuguesa (AO90). (1990). Retrieved from http://www.portaldalinguaportuguesa.org/acordo.php?action=acordo&version=1990

Formulário Ortográfico de 1943 (FO43). (1943). Retrieved from http://www.portaldalinguaportuguesa.org/acordo.php?action=acordo&version=1943

Acordo Ortográfico de 1945 (AO45). (1945). http://www.portaldalinguaportuguesa.org/acordo.php?action=acordo&version=1945

Feltim, V.D. (2004). *Suporte computacional à escrita científica em português*. (Doctoral dissertation, ICMC - USP/São Carlos).

Branco, A., Rodrigues, J., Costa, F., Silva, J., & Vaz, R. (2014). Assessing automatic text classification for interactive language learning. In *Proceedings of iSociety2014 - IEEE International Conference on Information Society IEE* (pp. 72–80).

Vilela, M. (2002). As expressões idiomáticas na língua e no discurso. In I. Duarte, Barbosa, J., S. Matos, & T. Hüsgen (Orgs.), *Actas do encontro comemorativo dos 25 anos do CLUP* (pp. 160–189). Retrieved from https://repositorio-aberto.up.pt/bitstream/10216/18051/2/7146000079120.pdf

Appendix A

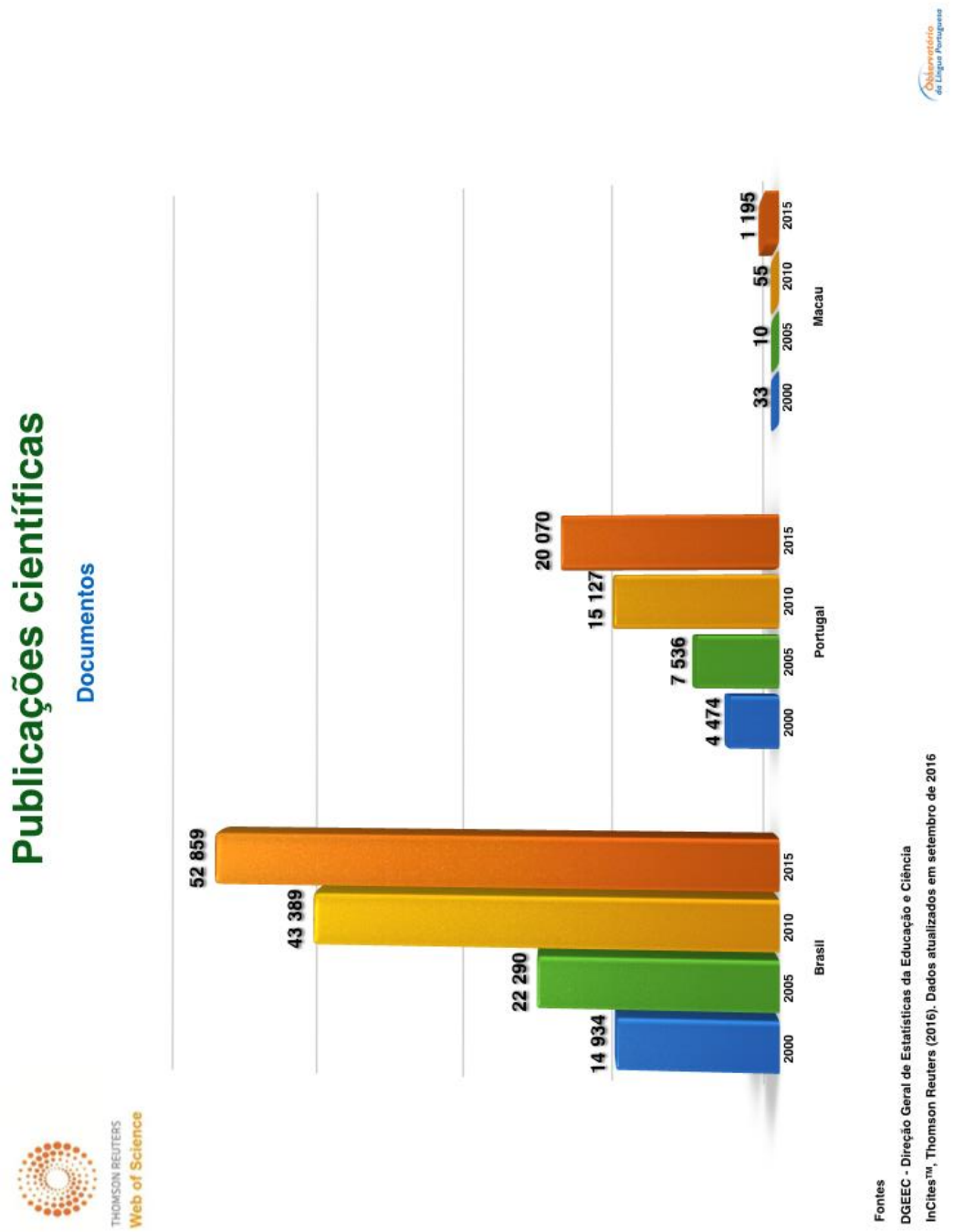


Figure 13 1 Number of scientific publications per year and per country (Observatório da Língua Portuguesa)

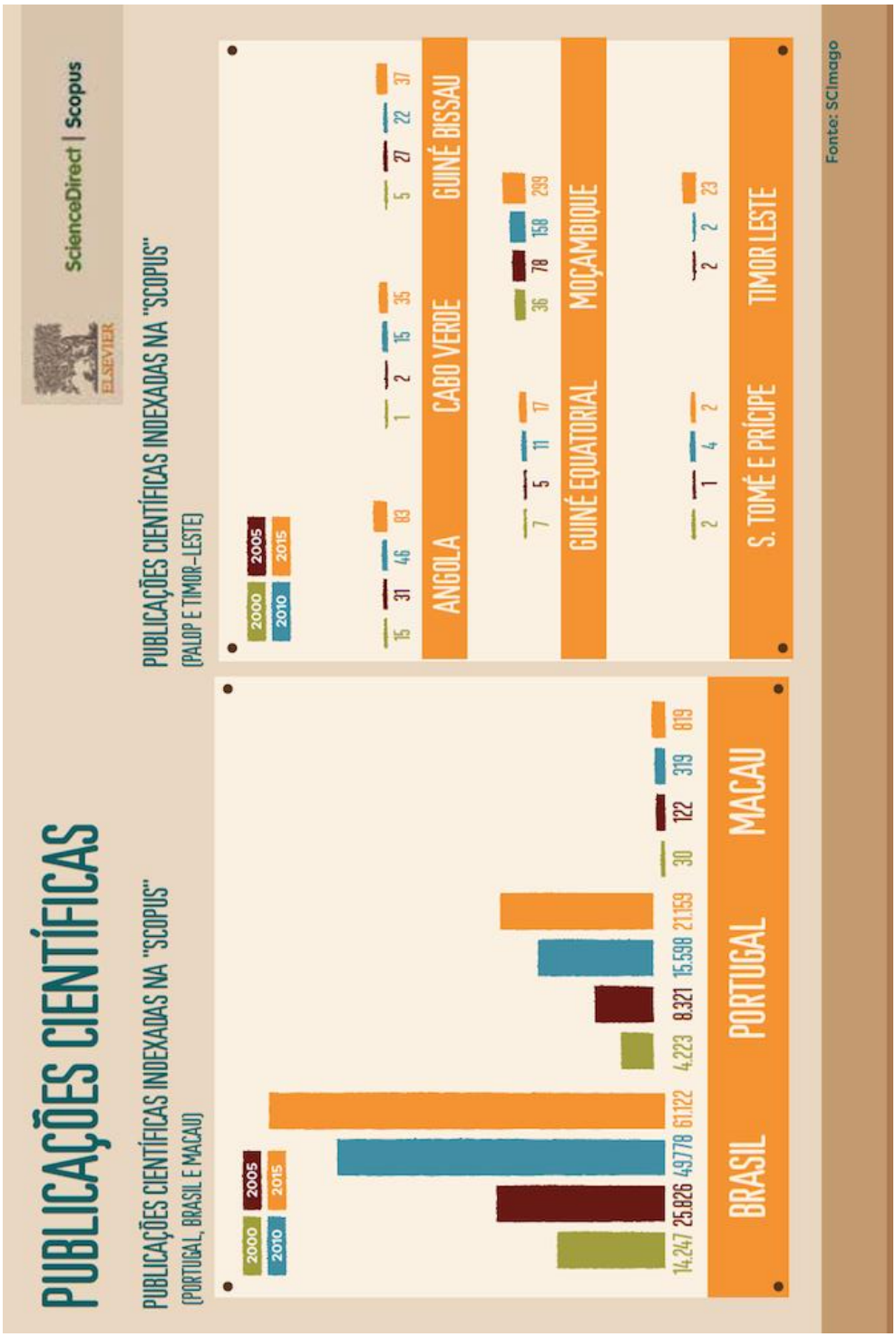


Figure 13 2 Scientific publications indexed in Scopus, per year and per country (Observatório da Língua Portuguesa)

Figure 13 3 Percentage of publication from CPPL countries, per language of publication (Observatório da Língua Portuguesa)

Appendix B

Figure 13 4 CEPRIL

Figure 13 5 CLUL

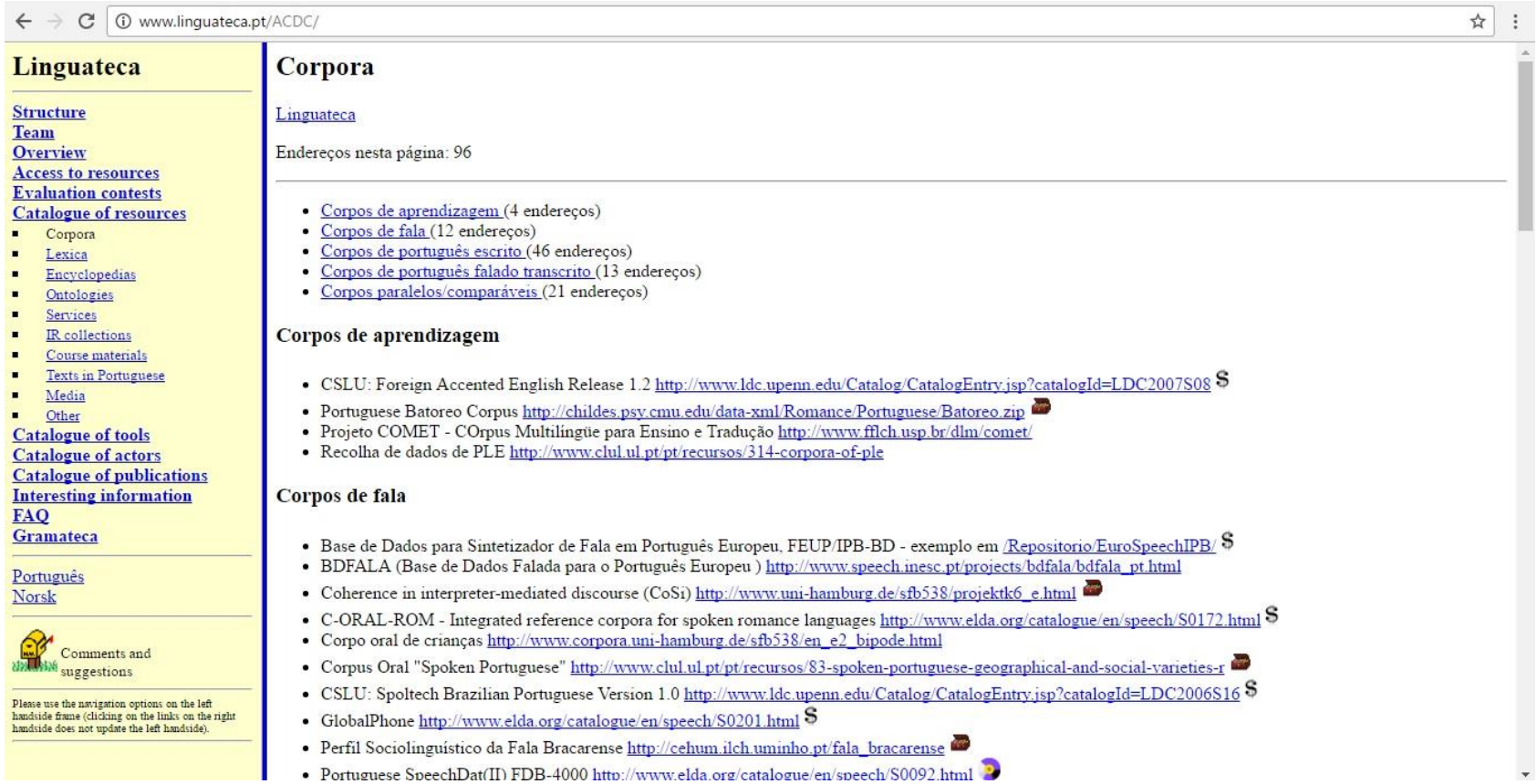


Figure 13 6 Linguateca

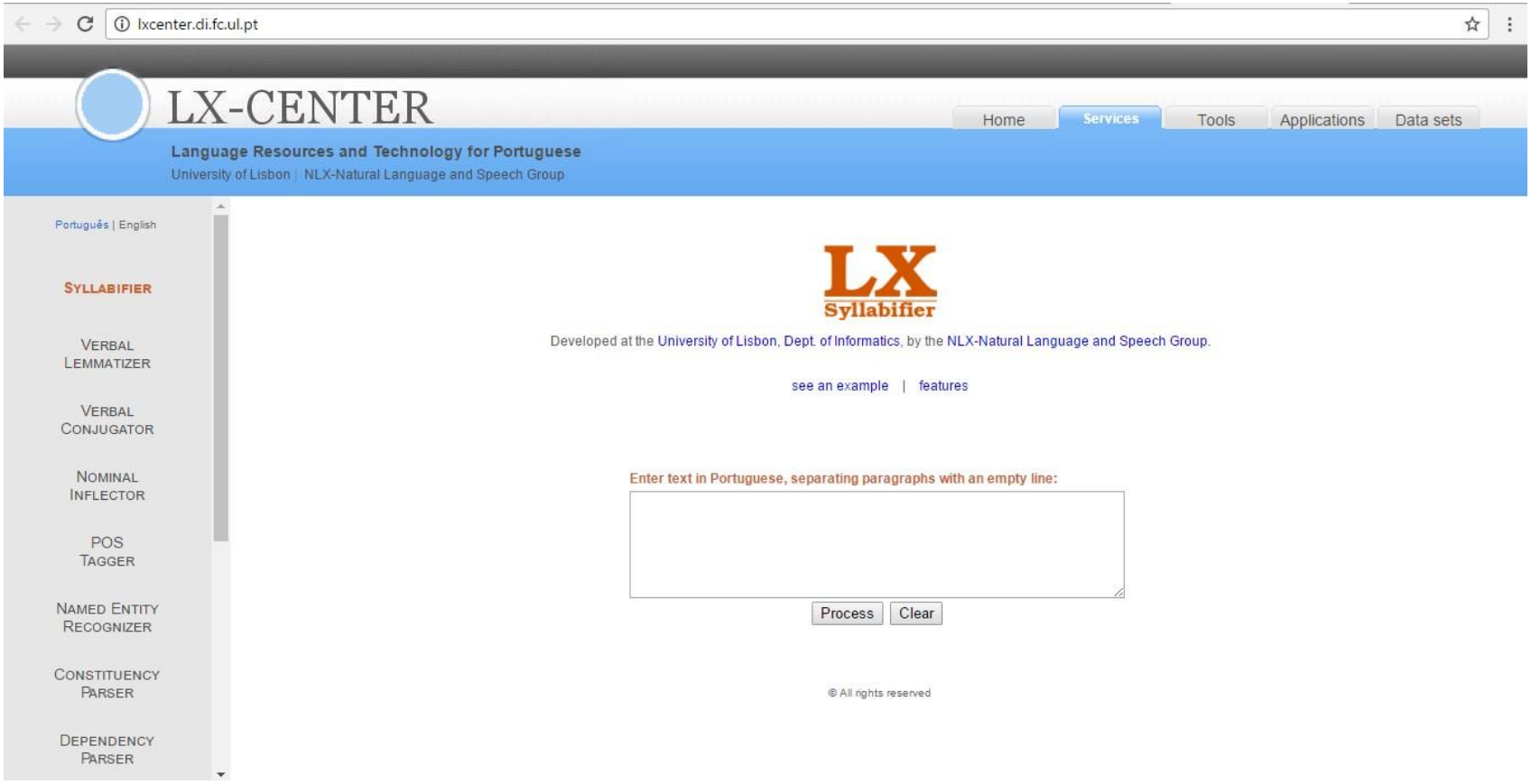


Figure 13 7 LX-Center

Figure 13 8 NILC

Appendix C

Tagset for Portuguese (pt)

Part of Speech: `adjective`

Position	Atribute	Values
0	category	**A**:*adjective*
1	type	**O**:*ordinal*; **Q**:*qualificative*; **P**:*possessive*
2	degree	**S**:*superlative*; **V**:*evaluative*
3	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*
4	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
5	possessorpers	**1**:*1*; **2**:*2*; **3**:*3*
6	possessornum	**S**:*singular*; **P**:*plural*; **N**:*invariable*

Part of Speech: `adposition`

Position	Atribute	Values
0	category	**S**:*adposition*
1	type	**P**:*preposition*

Part of Speech: `verb`

Position	Atribute	Values
0	category	**V**:*verb*
1	type	**M**:*main*; **A**:*auxiliary*; **S**:*semiauxiliary*
2	mood	**I**:*indicative*; **S**:*subjunctive*; **M**:*imperative*; **P**:*pastparticiple*; **G**:*gerund*; **N**:*infinitive*
3	tense	**P**:*present*; **I**:*imperfect*; **F**:*future*; **S**:*past*; **C**:*conditional*; **M**:*plusquamperfect*
4	person	**1**:*1*; **2**:*2*; **3**:*3*
5	num	**S**:*singular*; **P**:*plural*
6	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*

Part of Speech: `conjunction`

Position	Atribute	Values
0	category	**C**:*conjunction*
1	type	**C**:*coordinating*; **S**:*subordinating*

Part of Speech: `determiner`

Position	Atribute	Values
0	category	**D**:*determiner*
1	type	**A**:*article*; **D**:*demonstrative*; **E**:*exclamative*; **I**:*indefinite*; **T**:*interrogative*; **N**:*numeral*; **P**:*possessive*
2	person	**1**:*1*; **2**:*2*; **3**:*3*
3	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*
4	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
5	possessornum	**S**:*singular*; **P**:*plural*

Part of Speech: `noun`

Position	Atribute	Values
0	category	**N**:*noun*
1	type	**C**:*common*; **P**:*proper*
2	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*
3	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
4	neclass	**S**:*person*; **G**:*location*; **O**:*organization*; **V**:*other*
5	nesubclass	*Not used*
6	degree	**A**:*augmentative*; **D**:*diminutive*

Part of Speech: `number`

Position	Atribute	Values
0	category	**Z**:*number*
1	type	**d**:*partitive*; **m**:*currency*; **p**:*ratio*; **u**:*unit*

Part of Speech: `date`

Position	Atribute	Values
0	category	**W**:*date*

Part of Speech: `interjection`

Position	Atribute	Values
0	category	**I**:*interjection*

Part of Speech: `pronoun`

Position	Atribute	Values
0	category	**P**:*pronoun*
1	type	**D**:*demonstrative*; **E**:*exclamative*; **I**:*indefinite*; **T**:*interrogative*; **N**:*numeral*; **P**:*personal*; **R**:*relative*
2	person	**1**:*1*; **2**:*2*; **3**:*3*
3	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*
4	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
5	case	**N**:*nominative*; **A**:*accusative*; **D**:*dative*; **O**:*oblique*
6	polite	**P**:*yes*

Part of Speech: `adverb`

Position	Atribute	Values
0	category	**R**:*adverb*
1	type	**N**:*negative*; **G**:*general*

Non-positional tags

Part of Speech: `punctuation`

Tag	Attributes
Fd	**pos**:*punctuation*; **type**:*colon*
Fc	**pos**:*punctuation*; **type**:*comma*
Flt	**pos**:*punctuation*; **type**:*curlybracket*; **punctenclose**:*close*
Fla	**pos**:*punctuation*; **type**:*curlybracket*; **punctenclose**:*open*
Fs	**pos**:*punctuation*; **type**:*etc*
Fat	**pos**:*punctuation*; **type**:*exclamationmark*; **punctenclose**:*close*
Faa	**pos**:*punctuation*; **type**:*exclamationmark*; **punctenclose**:*open*
Fg	**pos**:*punctuation*; **type**:*hyphen*
Fz	**pos**:*punctuation*; **type**:*other*
Fpt	**pos**:*punctuation*; **type**:*parenthesis*; **punctenclose**:*close*
Fpa	**pos**:*punctuation*; **type**:*parenthesis*; **punctenclose**:*open*
Ft	**pos**:*punctuation*; **type**:*percentage*
Fp	**pos**:*punctuation*; **type**:*period*
Fit	**pos**:*punctuation*; **type**:*questionmark*; **punctenclose**:*close*
Fia	**pos**:*punctuation*; **type**:*questionmark*; **punctenclose**:*open*
Fe	**pos**:*punctuation*; **type**:*quotation*
Frc	**pos**:*punctuation*; **type**:*quotation*; **punctenclose**:*close*
Fra	**pos**:*punctuation*; **type**:*quotation*; **punctenclose**:*open*
Fx	**pos**:*punctuation*; **type**:*semicolon*
Fh	**pos**:*punctuation*; **type**:*slash*
Fct	**pos**:*punctuation*; **type**:*squarebracket*; **punctenclose**:*close*
Fca	**pos**:*punctuation*; **type**:*squarebracket*; **punctenclose**:*open*

Figure 13 9 Tagset for Portuguese Freeling v3

Appendix E

Table 13.2 Word sketch corpus query language (CQL) searches

[ws("estudo-n","e_ou","trabalho-n")]

[ws("apresentar-v","e_ou","discutir-v")]

[ws("grande-j","e_ou","importante-j")]

[ws("ainda-r","e_ou","já-r")]

[ws("estudo-n","Adj-Part mod %w_N","presente-j")]

[ws("momento-n","Adj-Part mod %w_N","determinar-v")]

[ws("grande-j","%w_Adj-Part mod N","parte-n")]

[ws("determinar-v","%w_Adj-Part mod N","momento-n")]

[ws("estudo-n","%w_N mod por Adj-Part","recente-j")]

[ws("estudo-n","%w_N mod por Adj-Part","realizar-v")]

[ws("realizar-v","N mod por %w_Adj-Part","estudo-n")]

[ws("grande-j","N mod por %w_Adj-Part","número-n")]

[ws("estudo-n","%w_N ser-estar N","parte-n")]

[ws("atividade-n","N ser-estar %w_N","trabalho-n")]

[ws("estudo-n","%w_N ser-estar Adj","necessário-j")]

[ws("grande-j","N ser-estar %w_Adj","diferença-n")]

[ws("estudo-n","%w_N suj de V","mostrar-v")]

[ws("apresentar-v","N suj de %w_V", "grupo-n")]

[ws("ainda-r","%w_Adv mod Adj-Part","maior-j")]

[ws("já-r","%w_Adv mod Adj-Part","realizar-v")]

[ws("grande-j","Adv mod %w_Adj-Part","tão-r")]

[ws("obter-v","Adv mod %w_Adj-Part","assim-r")]

[ws("grande-j","%w_Adj-Part mod por Adv","demais-r")]

```
[ws("obter-v","%w_Adj-Part mod por Adv","através-r")]

[ws("através-r","Adj-Part mod por %w_Adv","adquirir-v")]

[ws("apresentar-v","%w_V mod por Adv", "como-r")]

[ws("ainda-r","V mod por %w_Adv","referir-v")]

[ws("ainda-r","%w_Adv mod V","permanecer-v")]

[ws("apresentar-v","Adv mod %w_V","aqui-r")]

[ws("estudo-n","%w.*N","caso-n")]

[ws("apresentar-v","%w.*N", "artigo-n")]

[ws("estudo-n","\.\.\..*%w","objeto-n")]

[ws("estudo-n","\.\.\..*%w", "participar-v")]

[ws("apresentar-v","%w_V obj N", "resultado-n")]

[ws("estudo-n","V obj %w_N", "realizar-v")]

[ws("tender-v","\.\.\.\%w.*Vinf", "aumentar-v")]

[ws("necessidade-n","\.\.\.\%w.*Vinf", "desenvolver-v")]

[ws("capaz-j","\.\.\.\%w.*Vinf", "produzir-v")]

#[ws("apresentar-v",%w_verbo com se \+ N, "resultado-n")]

[ws("estudo-n","verbo com se \+ %w_N", "realizar-v")]

[ws("realizar-v","passiva impessoal","estudo-n" )]

[ws("estudo-n","sujeito da passiva impessoal", "realizar-v")]

[ws("estudo-n","sujeito da passiva pessoal", "aprovar-v")]

[ws("aprovar-v","passiva pessoal", "estudo-n")
```